Journal of Educational Psychology

Wayne H. Holtzman, Editor
University of Texas

Advisory Editors

Robert H. Beezer
United States Office of Education

Daniel E. Berlyne
University of Toronto

C. Victor Bunderson
University of Texas

Lee J. Cronbach
Stanford University

Nathaniel Gage
Stanford University

Robert M. Gagné
University of California, Berkeley

Robert Glaser
University of Pittsburgh

Ernest R. Hilgard
Stanford University

Philip W. Jackson
University of Chicago

Arthur R. Jensen
University of California, Berkeley

Herbert J. Klausmeier
University of Wisconsin

Nathan Kogan
Educational Testing Service, Princeton

Arthur A. Lumsdaine
University of Washington

Carson McGuire
University of Texas

Paul Mussen
University of California, Berkeley

C. Robert Pace
University of California, Los Angeles

David G. Ryans
University of Hawaii

Robert L. Thorndike
Teachers College, Columbia University

Volume 59, 1968

Helen Orr
Managing Editor

165

Henry S. Salmon
Advertising and Business Manager

Barbara Virgo
Technical Editor

Robert D. Gorman
Subscription Manager

PUBLISHED BIMONTHLY BY:
THE AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.
1200 Seventeenth Street, N. W.
WASHINGTON, D. C. 20036

$10.50

Copyright, 1968, by the American Psychological Association, Inc.

Bureau of Educl. & Psyl. Research
S.C.E.R.T.
Date 6-3-81
Acc. 5679

CONTENTS OF VOLUME 59

Monograph Supplements

JOURNAL OF
EDUCATIONAL PSYCHOLOGY

Volume 59 February 1968 Number 1, Part 1 of 2 Parts

CONTENTS

© 1968 by the American Psychological Association, Inc.

Journal of Educational Psychology
1968, Vol. 59, No. 1, 1–5

INFLUENCE OF PUPILS' ATTITUDES ON PERCEPTION OF TEACHERS' BEHAVIORS AND ON CONSEQUENT SCHOOL WORK[1]

JANICE B. GOLDBERG
University of Maryland, Baltimore

LIBRARY

In a study to determine whether attitudes toward authority and school work are associated with differential perception of teachers' behaviors and school performance, 254 8th- and 9th-grade boys classified as high or low on the California F Scale, Flexibility Scale, and Compulsivity Scale rated teacher behavior and reported the amount of school work performed. High compulsives perceived teachers as significantly less authoritarian than did low compulsives and did less work when the teacher was perceived as nonauthoritarian.

Pupils who achieve well in one teacher's classroom may achieve poorly in another teacher's class. Similarly, in any classroom some pupils achieve well, while others, equally intelligent, achieve poorly. Little is known about *differential* pupil reaction to teachers and its consequent effect upon pupils' school performance.

Some investigations of pupil-teacher relationships have used pupils' ratings of teachers' classroom behaviors which are conceptualized along personality dimensions related to the authoritarian versus the nonauthoritarian personality pattern (Amidon & Flanders, 1961; Cogan, 1954; Flanders, 1951). The intent of these studies has been to show that authoritarian-related teacher behaviors elicit pupil anxiety which results in lowered pupil achievement. The findings of these studies are inconclusive. The validity of using pupils' ratings as a method for determining differential pupil reaction to teachers stems from the fact that pupils observe more of the teacher's typical behavior than is usually available to the outside observer. Moreover, pupils are directly involved in the teaching-learning process.

There are two major shortcomings in most pupil rating studies. One is the pooling of all pupils' ratings without consideration of individual differences in pupils' perceptions despite the fact that extensive research has shown that individual personality factors influence perception (Bruner, 1958). The other is the use of broad variables, for example, "liking the teacher." Such global variables do little to clarify the complexity of teacher-pupil relationships and require considerable inference on the part of pupils. To avoid these shortcomings this study investigated the relationships between pupils differentiated in their attitudes toward authority and their perceptions of specific, denotable teacher behaviors. These specific behaviors require less inference by pupils.

The study rests on the premise that differential pupil reaction to teachers may be due to underlying attitudinal factors which influence pupils' perceptions of teachers' behaviors and their consequent performance of school work. It is also assumed that teachers' attitudinal factors predispose them to behave in a particular fashion and that these behaviors can be identified.

While many personality factors influence perception and behavior, this study relies upon the measurement of attitudinal variables comprising the authoritarian versus the nonauthoritarian personality dimension (Adorno, Frenkel-Brunswik, Levinson, & Sanford, 1950) as a significant determinant of teacher behavior and as an influence on pupil perception. In view of findings that high and low authoritarians

[1] This study is based on a dissertation submitted to the Graduate School of Education of Harvard University in partial fulfillment of the requirements for the Doctor of Education degree. The research was supported by a grant from the Milton Fund. The author wishes to thank her advisors, D. W. Oliver and G. W. Goethals, for their assistance in the execution of the study.

differ in their perceptions of others (Scodel & Mussen, 1953), in their study habits (Gladstein, 1957), and in their ability to learn different kinds of subject matter (Neel, 1959), it is suggested that pupils who are differentiated in their attitudes toward authority will differ in their perceptions of teachers' behaviors related to the authoritarian versus the nonauthoritarian personality dimension (Adorno et al., 1950).

These hypotheses were tested: (*a*) Pupils differentiated as high or low on the California F Scale, on the Flexibility Scale, and on the Compulsivity Scale will differ in their perceptions of teachers' classroom behaviors. (*b*) The ratings of teachers' classroom behaviors by pupils differentiated as high or low on the three attitude scales are related to the amount of required and self-initiated work pupils report they perform.

METHOD

There are three important features in the research design. One is the use of specific, denotable classroom teacher behaviors which are operationally defined in terms of the variables underlying the California F Scale (Adorno et al., 1950) rather than the more common use of global variables to describe teachers' behaviors. The characterization of teachers' behaviors is based on a well-defined personality theory, that is, the authoritarian personality dimension, since the variables comprising this dimension are expressed by teachers in many ways in the classroom. The second is the assessment of pupils' attitudes on a personality dimension which taps attitudes toward authority. This seems appropriate since the authority vested in the teacher is an important construct in the pupil's daily school life. The third important feature is the use of two unique criterion variables devised by Cogan (1954)—the amount of required work and self-initiated work performed by pupils. While these criteria do not directly measure pupil change, for example, as measured by standardized achievement tests, they avoid the pitfalls of these tests. Cogan argued that performance of pupil *work* is closely related to pupil *change* (or gain) and "intervenes just prior to pupil change."

Subjects

Subjects were 254 eighth- and ninth-grade boys and their 12 male social studies teachers in three junior high schools in a Boston suburb.

Instruments

All measures (pupils' attitudes, pupils' ratings of teacher behaviors, and pupils' reports of required and self-initiated work) were secured from one questionnaire. Administration procedures were standardized, and pupils were given assurances of anonymity.

Independent Variables

To differentiate pupils according to their attitudes toward authority, three instruments were used: (*a*) McGee's (1955) revised version of the California F Scale was modified for this study. High scorers (authoritarians) tend to be intolerant of ambiguity, to be rigid in their thinking, and to show perceptual distortion when rating others. Split-half reliability is .86. (*b*) Gough's (1956) Flexibility Scale was also modified for this study. This scale measures desire for order and certainty, especially in intellectual matters. High scorers resist learning ambiguous material found in social studies content, a resistance associated with the authoritarian's intolerance of ambiguity. The scale correlates ($r = .59$, $p < .01$) with the F Scale described above. Inter-item reliability is .75. (*c*) Berlak's (1959) Compulsivity Scale was shortened for this study. This scale measures attention to detail and order, particularly in school work. High scorers (high compulsives) may be said to have a strong desire to do well in school while the converse is true for low compulsives. These assumptions regarding the differences in the attitudes of high and low compulsive pupils seem tenable in view of Oliver and Shaver's (1962) finding of a strong relationship ($r = .57$, $p < .005$) between Compulsivity and "need cognition," defined as "a desire to do well in school." The overall rigidity of behavior measured by the scale reflects attitudes consistent with authoritarianism. The scale correlates ($r = .20$, $p < .05$) with the F Scale. Inter-item reliability is .76. A Likert-type scale is provided for each of the three attitude measures with six possible responses to each item. The responses range from "strong agreement" to "strong disagreement."

Dependent Variables

Descriptive teacher behaviors were related, by hypothesis, to variables representing the underlying personality trends measured by the California F Scale. In turn, these descriptive behaviors, hypothesized to be manifestations of the authoritarian versus the nonauthoritarian personality pattern, were matched with *specific, denotable teacher behavior items* drawn from Cogan's (1954) Pupil Survey. The criterion guiding the process of determining which *denotable behavior item* corresponded to the *descriptive* behavior was the functional relevance of the item to the behavior. The process of matching descriptive behaviors and thereby relating F-Scale variables to denotable teacher behaviors is demonstrated as follows. One F-Scale variable comprising the authoritarian personality structure is termed "Anti-intraception." This personality trend is characterized by impatience with the subjective and the tender-minded. It was hypothesized that such a teacher would

tend to be unconcerned with what pupils think and feel, and might be contemptuous of the academically poor pupil. The descriptive behavioral manifestation of this variable is the statement, "Teacher is unsympathetic with a pupil's failure at a task." The operational definition of this behavior and, therefore, of the F-Scale variable is the questionnaire item: "When we give a wrong answer in class, our teacher says we are 'slow,' 'not smart,' etc."

The 25 specific authoritarian teacher behavior items characterize the authoritarian teacher, in part, as strongly directive, impatient with academically inferior pupils, and generally rejecting of pupils. High scores on these items represent the extent to which pupils perceive teachers as authoritarian.

A similar procedure was used to relate descriptive nonauthoritarian teacher behaviors to specific Pupil Survey items. The 23 specific nonauthoritarian teacher behaviors represent personality trends hypothesized to be *opposite* to the meaning of the F-Scale variables.[2] These items characterize the nonauthoritarian teacher, in part, as permissive, more concerned with individual pupils' needs, and generally accepting of pupils. High scores on these items represent the extent to which pupils perceive teachers as nonauthoritarian.

A 5-point frequency scale was provided for each of the items. Responses to the items range from "Almost never" to "Very often." Cogan (1954) reports a reliability coefficient of .96 for these teacher behavior items. Pupils' ratings of these items are the dependent variables of the study.

Criterion Variables

To determine the amount of school work performed by pupils, two measures were used: (*a*) 24 items concerned with homework represent the amount of *required work* a pupil does, and (*b*) 21 items represent the extent to which a pupil does extra, unassigned (*self-initiated*) school work. Some examples are:

Required work: "Give a report"

Self-initiated work: "I make extra graphs, charts, etc."

The items were rated on a 6-point frequency scale. Responses to the items range from "Almost never" to "Almost always." The reliability coefficient for the required work items is .94 and .89 for the self-initiated work items (Cogan, 1954).

RESULTS AND DISCUSSION

Hypothesis 1 is partially confirmed. Table 1 shows that compulsivity in the total sample is strongly related to pupils' perceptions of teachers' behaviors. Pupils' F-Scale and Flexibility Scale scores are not related to their perceptions of teachers' behaviors.

While compulsivity is a component of the F Scale ($r = .20$, $p < .05$) and of the Flexibility Scale ($r = .27$, $p < .005$), its influence on pupil perception of teacher behavior may be due to the fact that it measures school-related attitudes rather than the generalized attitudes measured by the F Scale and the Flexibility Scale.

The t tests in Table 2 show that when pupils are differentiated as high or low on the Compulsivity Scale their ratings of teachers' behaviors are significantly different. High compulsives, those who work carefully in order to do well in school, perceive teachers as more nonauthoritarian. Low compulsives, those who may be less concerned with doing well in school, perceive teachers as more authoritarian.

TABLE 1

CORRELATIONS BETWEEN PUPILS' SCORES ON THE ATTITUDE SCALES AND THEIR PERCEPTIONS OF TEACHERS' BEHAVIORS AND THE AMOUNT OF REQUIRED AND SELF-INITIATED WORK PERFORMED

	Pupils' scores		
	F Scale	Flexibility Scale	Compulsivity Scale
Teachers' behaviors			
Authoritarian	−0.07	−0.08	−0.23*
Nonauthoritarian	0.04	0.08	0.23*
Work performed			
Required	0.02	0.08	0.35**
Self-initiated	0.01	0.02	0.43**

Note.—$N = 254$.
* $p < .01$.
** $p < .005$.

TABLE 2

MEANS OF HIGH AND LOW COMPULSIVE PUPILS' RATINGS OF TEACHERS' BEHAVIORS

Teachers' behaviors	Mean pupil ratings			
	Low-compulsive	High compulsive	*df*	*t*
Authoritarian	57.21	51.99	252	2.29*
Nonauthoritarian	80.16	84.97	252	−2.51**

* $p < .05$.
** $p < .01$.

[2] Complete tables of descriptive teacher behaviors and specific items may be found in the author's unpublished doctoral dissertation (Goldberg, 1965).

The differences in perception by the two groups of pupils may lie in differential treatment of these pupils. The authoritarian teacher is characterized, in part, as strongly directive, impatient with academically poor pupils, and may be unconcerned with pupils' personal needs and goals. The authoritarian teacher tends to demand good school performance, to insist on strict order, and to conform to a rigid routine.

Since the Compulsivity Scale measures attention to detail and order in school work, it has been assumed that the high compulsive wants to do well in school. He is probably the academically strong pupil who conforms to teacher expectations of good work. It may be that less demand is made of him by his teacher since he does good work in accordance with his own inner needs. He may be favorably treated in that he gets less criticism from his teacher. Thus, he perceives the teacher as more nonauthoritarian. The low compulsive may care less about good school performance. He is probably the academically weak pupil who does not meet teacher expectations of good work. The teacher may excessively criticize him and may persistently require him to do more careful work. Thus, the low-compulsive pupil perceives the teacher as more authoritarian. It is also possible that pupils' attitudes toward the teacher as an authority figure may result in perceptual distortion of the teacher's behavior. The high compulsive, with his need to do well in school, may perceive teacher demands for good work as aiding him to do well in school and, therefore, as reasonable behavior. Thus, he rates the teacher as less authoritarian. The low compulsive, having less need to do well in school, may perceive this kind of teacher behavior as unreasonable and he rates the teacher as more authoritarian.

Hypothesis 2 is partially confirmed. Table 1 also shows that compulsivity in the total sample is positively related to the amount of work pupils perform. However, no significant relationship exists between either F-Scale or Flexibility Scale scores and amount of school work performed.

Two-way analyses of variance show differences in pupils' compulsivity and differences in pupils' perceptions of teachers' nonauthoritarian behaviors do influence their performance of required and self-initiated work. There is a highly significant interaction effect between pupils' compulsivity and pupils' perceptions of nonauthoritarian teacher behaviors with self-initiated work as the criterion measure ($F = 19.17$, $df\ 1/218$, $p < .01$). The required F ratio at this level is 6.76. Interaction between pupils' compulsivity and their perceptions of nonauthoritarian teacher behaviors with required work as the criterion measure is significant at the .05 level.

The cell means for these analyses show that high-compulsive pupils do *less* work when the teacher is perceived as nonauthoritarian. Low-compulsive pupils do *more* work when the teacher is perceived as nonauthoritarian. Although the data derived for Hypothesis 1 indicate that high-compulsive pupils *perceive* teachers as more nonauthoritarian, these same cell means show that this perception of the teacher influences their performance of school work. These findings suggest that perception of teacher behavior as nonauthoritarian may conflict with high-compulsive pupils' need for a directive, demanding teacher. Thus, perception of nonauthoritarian behavior appears to serve as cues for anxiety in high-compulsive pupils, resulting in less performance of school work. Conversely, perception of nonauthoritarian behavior seems to encourage low-compulsive pupils to do more work, probably because this kind of teacher tends to have warmer relationships with pupils—even those who have less desire to do well in school.

Two-way analyses of variance of pupils' compulsivity and their perceptions of teachers' authoritarian behaviors show no significant interaction effects. However, trends in the cell means reveal that high-compulsive pupils do more work when

they perceive the teacher as authoritarian than do low-compulsive pupils. This lends weight to suggestions made earlier that high-compulsive pupils perceive this kind of teacher behavior as enabling them to fulfill their need to do well in school. Low-compulsive pupils do less work when the teacher is perceived as authoritarian. Thus, it is possible that such teacher behaviors may serve as cues for anxiety resulting in lowered school performance for those pupils who are less concerned about good school work.

The findings tend to support the conclusion that pupils differentiated in their attitudes—in this case on a measure of compulsivity—do perceive different kinds of teachers' behaviors differently and that this differential in perception influences the consequent amount of school work performed.

In view of contemporary concern for teaching "disadvantaged" children who tend to have little interest in good school performance, these results may be helpful in selecting teachers for these children as well as for studying their learning patterns.

REFERENCES

ADORNO, T. W., FRENKEL-BRUNSWIK, E., LEVINSON, D. J., & SANFORD, R. N., *The authoritarian personality.* New York: Harper, 1950.

AMIDON, E., & FLANDERS, N. A. The effects of direct and indirect teacher influence on dependent-prone students learning geometry. *Journal of Educational Psychology,* 1961, **52,** 286–291.

BERLAK, H. Rigidity scale. Unpublished manuscript. Cambridge: Harvard Graduate School of Education, 1959.

BRUNER, J. S. Social psychology and perception. In E. E. Maccoby, T. R. Newcomb, & E. L. Hartley (Eds.), *Readings in social psychology.* (3rd ed.) New York: Holt, 1958. Pp. 85–94.

COGAN, M. L. The relation of the behavior of teachers to the productive behavior of their pupils. Unpublished doctoral dissertation, Harvard University, 1954.

FLANDERS, N. A. Personal-social anxiety as a factor in experimental learning situations. *Journal of Educational Research,* 1951, **45,** 100–110.

GLADSTEIN, G. A. The relationship between study behavior and personality for academically successful students. Unpublished doctoral dissertation, University of Chicago, 1957.

GOLDBERG, J. B. The influence of pupils' attitudes on their perceptions of teachers' behaviors and on their consequent performance of school work. Unpublished doctoral dissertation, Harvard University, 1965.

GOUGH, H. G. *Manual for the California Psychological Inventory.* Palo Alto: Consulting Psychologists Press, 1956.

MCGEE, H. M. Measurement of authoritarianism and its relation to teachers' classroom behavior. *Genetic Psychology Monographs,* 1955, **52,** 89–146.

NEEL, A. F. The relationship of authoritarian personality to learning: F Scale scores compared to classroom performance. *Journal of Educational Psychology,* 1959, **50,** 195–199.

OLIVER, D. W., & SHAVER, J. P. *The analysis of public controversy: A study in citizenship education.* (Report from the Laboratory for Research in Instruction.) Cambridge: Harvard Graduate School of Education, 1962.

SCODEL, A., & MUSSEN, P. Social perceptions of authoritarians and non-authoritarians. *Journal of Abnormal and Social Psychology,* 1953, **48,** 181–184.

(Received December 16, 1966)

Journal of Educational Psychology
1968, Vol. 59, No. 1, 6-11

EFFECT OF MASSED PRACTICE ON THE COMPREHENSION OF TIME-COMPRESSED SPEECH[1]

DAVID B. ORR[2] AND HERBERT L. FRIEDMAN
American Institutes for Research, Washington, D. C.

The experiment described tested the effectiveness of the time-compressed speech technique (tape recordings presented at rates exceeding recording rates, but with normal pitch) in presenting material under conditions of massed practice in listening to time-compressed speech. A small group of Ss received practice material for about 7 hr. a day for 5 consecutive days at rates about 2½ times normal speaking rates (at 425 words per minute). Benchmark passages and tests were presented daily. Results showed the comprehension increased from a mean of about 40% of normal speed comprehension on Day 1 to a mean of 70% comprehension in Day 5. While effective, the massed-practice procedure produced no better performance in a total of 35 hours of practice than previous experiments using spaced practice of 1–2 hr. per day produced in a total of 12–15 hr. of practice.

For some years there has been a growing interest in the extent to which the human being can comprehend auditory material presented at rates more rapid than normal. Early work in this area was frustrated by the fact that speeding up recorded material produced a frequency or pitch shift which, in itself, interfered with comprehension as much as did the rapidity of the presentation. However, beginning about 15 years ago, interest began to develop in a process of time-compression which would preserve the normality of pitch and yet, at the same time, permit the speed-up of the presentation. The first work along these lines was done by Garvey (1953), who carefully sliced out systematic deletions from a piece of recording tape, splicing the remaining pieces together to form a continuous record. Garvey's experimental results showed that subjects (*S*s) could understand his tape almost as well as they could one presented in a normal fashion. Shortly thereafter Fairbanks, Everett, and Jaeger (1954) developed an electro-mechanical device at the University of Illinois, which, in effect, performed the same thing automatically that Garvey had done with a razor blade. Again, the results were the same. Fairbanks, Guttman, and Miron (1957) reported that *S*s were able to comprehend this material at rates ranging close to 300 words per minute with essentially no loss in comprehension. Somewhat later, Bixler, Foulke, Amster, and Nolan (1961) reported an attempt to apply the time-compression principle to the problem of providing material at a more rapid rate for blind students. Foulke and his group used a German machine called the Tempo Regulator which operated on a principle essentially similar to that used by Fairbanks. Again, results were favorable in that his *S*s reported being able to understand material at somewhat greater than normal presentation rates without much loss in comprehension.

Beginning in 1963 the senior author began to study the applicability of time-compression as a more general educational technique. Two basic questions were posed: (*a*) Can college-level students understand college-level material without loss of comprehension, when presented at modest rates of compression; and (*b*) as higher rates are introduced will the corresponding loss of comprehension be amenable to simple training (practice) routines? Initial experimentation by the present authors over a period of a year and half provided an affirmative answer to both of these questions (Orr, Friedman, & Williams, 1965).

[1] The research reported herein was supported by the United States Office of Education under Grant No. 7-48-7670-267. Acknowledgement of the assistance and interest of the monitor, Joseph A. Murnin, is gratefully made.

[2] Now with Software Systems, Inc., Falls Church, Virginia.

PROBLEM

Since that time the present authors have been engaged in experimental work designed to explore some of the other interesting questions which can be posed about the comprehension of time-compressed speech. Our initial experimentation had employed both graduated levels of practice ranging from approximately 325 words per minute up to 475 words per minute, and concentrated practice where all practice material was presented at approximately 425 words per minute. The results of an experiment conducted recently to determine the effectiveness of massive practice in learning to comprehend time-compressed speech are reported here.

In the case of both graduated and concentrated practice groups, practice sessions were confined to about an hour to an hour and a half per day and the entire period of experimentation was spread out over a period of 4–5 weeks. The purpose of the presently reported experimentation was to determine whether or not a concentration of massive practice in terms of a large number of clock hours per day could produce essentially comparable results in a few days. The rationale for the problem is two-fold. In the first place, if it should, as seems likely, become feasible to apply time-compressed speech as a general educational technique, it might be necessary to have naïve students spend some amount of time practicing the comprehension of time-compressed speech as a precursor to their regular studies. If this were the case, it would be desirable to have such a training course occupy a minimum number of days at the beginning of the term. Secondly, the experience of the Armed Services in recent years in attempting to teach a second language has shown a fair amount of success for intensive or immersion exposure to the target language.

PROCEDURE

The "immersion" study *Ss* consisted of seven male students, between ages 19 and 20, at the freshman or sophomore college level. English was their native language and none had a marked regional accent. The average letter grade for all students in their last semester in college was a C+. Two of the *Ss* had some training in rapid reading but in neither case was the course completed. None of the *Ss* had had any form of training in listening. All of the *Ss* were screened audiometrically for normal hearing.

In the first session, *Ss* were given a brief talk explaining that the purpose of this study was to provide intensive exposure to speeded speech, and to measure listening performance with periodic benchmark tests. They were also given a biographical data sheet to fill out which called for basic information about their backgrounds. They were then given alternate forms of the Nelson-Denny reading test, which measures reading comprehension, rate, and vocabulary. This was followed by the presentation of a historical passage (taken from the same book as the later benchmark passages) which was presented at normal recording speed (175 wpm). A standard multiple-choice test on the information contained in the passage was then given. A similar passage and test was then presented at 475 wpm as an initial measure of high-speed performance. The *Ss* were then asked to return for five consecutive weekdays, beginning on a Monday, from 9:00 AM to 9:00 PM.

During the next week, 12 novels were played at 425 wpm as practice materials for the *Ss*. The experiment was conducted in a semi-soundproofed room and materials previously compressed on the Tempo Regulator were played back on a Magnecord tape recorder through a Bogen amplifier and two Electro-voice speakers.

On each day listening material was presented for approximately 48 minutes without interruption. At the end of that time a brief written quiz, including both short-answer and essay questions, was administered to *Ss* during a 10-minute period. This was followed by a 5-minute rest period, after which the cycle was repeated. The *Ss* were given 1 hour for lunch during the afternoon, and 1 hour for dinner in the evening. During the latter part of each of the 5 days of exposure, a new benchmark passage and test, similar to the preexperimental material, was administered. Each test was presented at 425 wpm. Near the end of the fifth day, the initial high-speed benchmark passage was presented again at 475 wpm. The time schedule for the study is shown in Table 1.

At the end of testing on Wednesday and again in a postexperimental session the following Thursday, *Ss* were asked to rate the novels they had listened to, on a 5-point scale, covering the following aspects of the presentations: Overall ability to comprehend, intelligibility (clarity of individual words), difficulty of subject matter, interest in the book, quality of speaker's voice, and quality of speaker's diction. An alternate form of the Nelson-Denny reading test was presented as a postexperimental measure of change. The *Ss* were then given extensive debriefing questionnaires to complete calling for subjective comments on the

TABLE 1
PRACTICE MATERIALS AND SCHEDULE OF PRESENTATION
(all at 425 wpm)

Material	Time (in minutes)
Day 1	
Cheaper by the Dozen	45
Quiz	10
Break	5
Cheaper by the Dozen	40
Quiz	10
Cheaper by the Dozen	65
Quiz	10
Quiz	10
Break (lunch)	60
Diary of a Young Girl	67
Quiz	5
Break	10
Diary of a Young Girl	59
Quiz	10
Break	15
Diary of a Young Girl	43
Quiz	5
Break (dinner)	60
Benchmark passage C–4	12
Test C–4	13
Diary of a Young Girl	48
Quiz	5
Break	10
I Owe Russia $1200	79
Quiz	8
Day 2	
I Owe Russia $1200	60
Break	10
I Owe Russia $1200	25
Quiz	5
Quiz	5
The "Miracle" New York Yankees	49
Quiz	5
The "Miracle" New York Yankees	15
Break (lunch)	65
The "Miracle" New York Yankees	32
Quiz	5
The "Miracle" New York Yankees	32
Quiz	4
Break	20
To Kill a Mockingbird	57
Quiz	5
Break	15
To Kill a Mockingbird	54
Quiz	6
Break	10
To Kill a Mockingbird	54
Quiz	3
Break (dinner)	60
Benchmark passage C–5	10
Test C–5	10
To Kill a Mockingbird	44
Quiz	3
Break	2
To Kill a Mockingbird	45
Quiz	5
Day 3	
The Excitement of Science	53
Break	14
The Excitement of Science	9
Quiz	7
The Excitement of Science	30
Break	13
The Excitement of Science	31
Quiz	5
Break (lunch)	65
The Forgotten Pioneer	65
Quiz	5
Break	15
Man Eaters of Kumaon	47
Quiz	7
Break	18
Man Eaters of Kumaon	52
Quiz	7
Break	5
Day 3—Continued	
Man Eaters of Kumaon	19
Break (dinner)	60
Benchmark passage C–6	8
Test C–6	10
Man Eaters of Kumaon	35
Quiz	5
Break	10
Man Eaters of Kumaon	48
Quiz	5
Questionnaire	15
Day 4	
How to Win Friends and Influence People	50
Quiz	5
Break	14
How to Win Friends and Influence People	47
Quiz	6
Break	13
How to Win Friends and Influence People	33
Break (lunch)	64
How to Win Friends and Influence People	16
Quiz	3
How to Win Friends and Influence People	31
Break	15
How to Win Friends and Influence People	16
Quiz	4
Run Silent, Run Deep	24
Break	11
Run Silent, Run Deep	54
Break	10
Run Silent, Run Deep	39
Quiz	6
Break (dinner)	70
Benchmark passage C–7	8
Test C–7	10
Run Silent, Run Deep	30
Break	15
Run Silent, Run Deep	50
Break	5
Run Silent, Run Deep	17
Quiz	15
Day 5	
Run Silent, Run Deep	45
Break	10
Run Silent, Run Deep	60
Quiz	10
Break	18
America's Race for the Moon	31
Break (lunch)	65
America's Race for the Moon	18
Quiz	6
America's Race for the Moon	31
Break	12
America's Race for the Moon	8
Quiz	5
Riders of the Purple Sage	29
Break	15
Riders of the Purple Sage	27
Quiz	4
Riders of the Purple Sage	18
Break	9
Riders of the Purple Sage	38
Quiz	6
Break (dinner)	70
Benchmark passage C–2	8
Test C–2	10
Benchmark passage C–3[a]	8
Test C–3[a]	11
Riders of the Purple Sage	20
Break	10
Riders of the Purple Sage	61
Quiz	5

[a] Repeat of preexperiment measure.

TABLE 2

BENCHMARK TEST SCORES CORRECTED FOR CHANCE AND PERCENTAGES OF NORMAL SPEED SCORES

Subject	Initial measures		Practice at 425 wpm					Final measures
	Normal (175 wpm)	High speed (475 wpm)	Day 1	Day 2	Day 3	Day 4	Day 5	Repeated high speed (475 wpm)
A								
Score	16.25	0.00	6.50	2.75	7.50	8.75	9.37	9.37
%	100.0	0.0	40.0	16.9	46.2	53.8	57.7	57.7
B								
Score	14.0	5.21	9.50	9.75	8.75	7.50	9.58	12.29
%	100.0	37.2	67.9	69.6	62.5	53.6	68.4	87.8
C								
Score	25.00	8.33	15.00	8.75	16.25	10.00	17.70	11.45
%	100.0	33.3	60.0	35.0	65.0	40.0	70.8	45.8
D								
Score	22.50	8.95	16.25	11.50	14.00	13.75	16.09	15.62
%	100.0	39.8	72.2	51.1	62.2	61.1	71.3	69.4
E								
Score	19.25	1.46	0.75	3.75	6.75	14.00	9.58	6.66
%	100.0	7.6	3.9	19.5	35.1	72.7	49.8	34.6
F								
Score	14.25	0.00	1.25	6.50	1.25	3.75	11.45	4.16
%	100.0	0.0	8.8	45.6	8.8	26.3	80.4	29.2
G								
Score	10.0	2.08	3.00	0.50	4.00	6.50	9.16	3.96
%	100.0	20.8	30.0	5.0	40.0	65.0	91.6	39.6
Mean								
Score	17.32	3.72	7.46	6.21	8.36	9.18	11.84	9.07
%	100.0	19.8	40.4	34.7	45.7	53.2	70.0	52.0

procedures, materials, and potential usefulness of compressed speech in the educational setting.

Upon the completion of the experiment, each *S* was paid $100.00 plus a bonus of $25.00 to the *S* who demonstrated the greatest proficiency on the benchmark tests.

RESULTS AND DISCUSSION

Results of the benchmark tests in terms of number of questions correct, based on 25 item tests, corrected for chance, are shown in Table 2. Also shown is percentage of normal speed performance, calculated separately for each individual based on his own performance at normal (175 wpm) speed. It may be noted that there is a progression of means from 40.4% on Day 1 to 70.0% on Day 5, which is reasonably steady with the exception of a dropback on Day 2. In addition to this improvement, mean performance on the repeated high-speed passage (475 wpm) also improved from 19.8% to 52.0%, which is a significant improvement at the 1% level and significantly greater than that of a control group from previous experimentation. The progression of means is shown graphically in Figure 1.

With the exception of the 475 wpm passage and test, the figures shown in Table 1 and Figure 1 are based upon different tests, and test passages are thus independent estimates of performance. Passages were taken from the same book of early English history, however, and tests were constructed to be equivalent according to item statistics derived from the same population of students.

As an illustration of the extent to which variables such as type of test and material and subject variability can affect the results, however, one may consider the results for the short-answer tests and essay tests on the practice materials. These were not intended to do more than motivate *S*s to listen to the practice materials, and it was not possible to standardize these measures. The results emphasized the extent to which the evaluation of

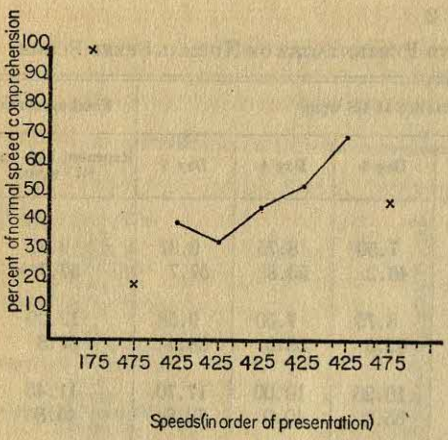


FIG. 1. Mean percentage of normal performance over 5 days ($N = 7$).

comprehension of compressed speech is dependent on such variables, showing no discernible trends and great subject variability. The lack of correlation between short answer and essay results ($r = 0$) indicated further that these measures are not very dependable. To the extent that research such as this depends upon such "pick-up" measures, it is certainly open to question.

With respect to pre- and posttest scores on the reading and standard listening tests, mean increases of about 7–8% were observed in each case. Of course, this modest figure was not significant with an N of only 7.

The results of the present experiment again confirm previous findings (Orr et al., 1965) that comprehension of time-compressed speech can be improved by simple practice routines to relatively high levels at speeds of about 2½ times normal. By the end of the week, all *S*s had reached the 50% comprehension mark, although several started from as low as 5–10% on the first day. There is reason to believe that some further increases would have been obtained had the experiment been prolonged. The effects of the training on other variables such as reading test scores and standard listening test scores, while in the right direction, were not great, however.

Another question of interest is the relative effectiveness of the method here employed to achieve approximately 70% comprehension at 425 wpm. During their 5 days of intensive exposure, these *S*s spent approximately 35 hours listening to compressed speech. Their results can be compared with those of three previous groups of similar composition, all of which received about 12–14 hours of practice distributed at about 1–2 hours per day, 2–3 days per week, over about 4–5 weeks.

Group A. Practice with speeds gradually increasing from 325 to 425 wpm;

Group B. Graduated practice from 325 to 425 wpm (with 3-minute breaks every 10 minutes);

Group C. High-speed practice (425 wpm only).

The mean result for Group A was 79% of normal at 425 wpm; for Group B, 80% at 425 wpm; and for Group C, 71% at 425 wpm. Thus, the investment of 12–14 hours of spaced practice produced results as good or better than the investment of almost three times as much practice in the present "immersion" study. A similar conclusion was reached after looking at the mean improvements from pre- to postexperimental scores on the repeated high-speed (475 wpm) test passage.

With respect to the subjective comments gathered on the debriefing questionnaire, a few comments can be made. All *S*s felt that practice had improved their ability to comprehend compressed speech and five of the seven felt that more practice would lead to further improvement. Attention wandering, particularly on less interesting parts of the practice material, was seen as one of the chief problems. However, most felt that their powers of concentration had been improved by the experiment. Finally, all *S*s felt that their performances would be improved with the use of learning aids, such as outlines, key words, abstracts, simultaneous availability of the text, and selected repetition. Further experimentation is being conducted.

The findings of the present study tend to reinforce the conclusions that time-compressed speech offers substantial possibilities as an educational technique.

Comparatively high levels of comprehension are possible after comparatively limited amounts of training at substantially increased rates of speed. Rather high levels of comprehension at 2½ times normal speed can be obtained in a period as short as one working week by means of concentrated practice. Thus, the findings do confirm the feasibility of a concentrated training course in compressed speech as a prelude to the regular school term, if the use of highly compressed speech should become a usual educational practice.

Of course, it must be emphasized that the application of time-compressed speech in the classroom does not mean the elimination of the professor. The technique is seen as a way to maximize his effectiveness. Routine material, survey material, etc., can be effectively presented in this manner, leaving the professor free to concentrate on those aspects of his teaching which demand more of his skill. For example, after presenting the basic materials for a day's discussion in half the usual time, he would be free to conduct a postpresentation critical analysis of the material, to organize round-table discussions, and to apply any one or more of the many educational techniques that the professor currently cannot apply in the classroom through a lack of time.

It is recognized that the proposed techniques are not likely to be suitable for all kinds of material, and currently experimentation is underway to determine the types of material for which the proposed technique is most suitable. Further, the technique is not equally effective with all students (as what technique is?), and inquiries are underway to attempt to determine those student characteristics which seem to interact most closely with success in comprehending time-compressed speech.

REFERENCES

BIXLER, R. H., FOULKE, E., AMSTER, C. H. & NOLAN, C. Y. *Comprehension of rapid speech by the blind.* Louisville: University of Louisville Press, 1961.

FAIRBANKS, G., EVERETT, W. L., & JAEGER, R. P. Method for time or frequency compression-expansion of speech. *Transactions of Institute of Radio Engineers—Professional Groups,* 1954, **AU-2,** 7–11.

FAIRBANKS, G., GUTTMAN, N., & MIRON, M. S. Effects of time compression upon the comprehension of connected speech. *Journal of Speech and Hearing Disorders,* 1957, **22,** 10–19.

GARVEY, W. D. The intelligibility of speeded speech. *Journal of Experimental Psychology,* 1953, **45,** 102–108.

ORR, D. B., FRIEDMAN, H. L., & WILLIAMS, J. C. C. Trainability of listening comprehension of speeded discourse. *Journal of Educational Psychology,* 1965, **56,** 148–156.

(Received January 6, 1966)

Journal of Educational Psychology
1968, Vol. 59, No. 1, 12-15

EFFECT OF WORD ASSOCIATIONS ON READING SPEED, RECALL, AND GUESSING BEHAVIOR ON TESTS

S. JAY SAMUELS[1]
University of Minnesota

A paragraph containing words with high-associative (HA) relationships should be read faster and with better recall than a similar paragraph containing words with low-associative (LA) relationships. Mean reading time for elementary school Ss in the HA condition was 43.82 sec. and 58.81 sec. in the LA condition ($p < .05$). The mean number of questions answered correctly for HA was 9.50 and 5.04 for LA ($p < .001$). When college Ss read the same paragraphs, the mean time was 35.26 sec. for HA and 38.26 for LA ($p < .01$). The mean number of questions answered correctly was 9.69 for HA and 6.87 for LA ($p < .001$). When required to guess the correct answer, control Ss chose significantly more often alternatives which contained words having HA relationships with words in the stem of the question. Results on reading speed are discussed in terms of the effect of word associations on perceptual factors in word recognition.

Taylor (1963) has indicated that when sentence meaning does not suggest the next word in a sentence, the reader must engage in more careful visual analysis of the next word than when sentence meaning does suggest the next word. In support of this point of view, Goodman (1965) found that children were able to read words in the context of a sentence which they were unable to read when presented alone. He also states that one of the reasons why words are misread, resulting in time-consuming regressions, is that sentence meaning may miscue the reader. In an investigation of word-recognition speed, Tulving and Gold (1963) found that as the amount of information in a sentence increased, the time required to recognize a target word decreased. Similarly, O'Neil (1953) and Rouse and Vernis (1963) demonstrated that when word associates are tachistoscopically exposed in succession, recognizing the first word aids in recognizing the second word, and the stronger the association between the words, the lower the recognition threshold.

Not only do word associations influence speed of word recognition, but they influence recall as well. Rosenberg (1965) demonstrated that more words designated as stimulus or response words were recalled after hearing a paragraph containing highly associated words than after hearing a similar paragraph containing words with lower associative relationships. Deese (1959) also found that more words can be recalled after exposure to a word list containing associated words than after exposure to a word list containing words which were not associatively related.

Because of these findings, it was predicted that a paragraph containing words with high-associative relationships would be read faster with better recall than a similar paragraph containing words with low-associative relationships. It was also predicted that when subjects (*S*s) are required to answer multiple-choice questions without having read the paragraph upon which the questions are based, they would choose alternatives on the basis of the strength of the associative relationship between words in the stem of the question and the response alternatives.

Experiment I

Method

Subjects. The *S*s were fifth and sixth graders from Minneapolis elementary schools who were randomly assigned to read either a high-association (HA) or low-association (LA) paragraph and to answer questions based on the paragraph. Thus, 28 *S*s were assigned to HA and 26 *S*s to LA paragraphs. Forty *S*s were assigned to a control condition in which they answered questions based on the readings but did not actually read a paragraph.

Materials. Two 150-word paragraphs developed by Rosenberg (1965) were used. Rosenberg selected stimulus and response words of considerable

[1] Appreciation is extended to Barbara Best for her valuable help on this study.

strength from the New Minnesota Norms (Palermo & Jenkins, 1964), and these words were used in the HA paragraph. Words such as MAN, CHEESE, CARPET, and MOON were designated as stimulus words, and from one to five of their stronger associates, designated as response words, were included in the HA paragraph.

The LA paragraph contained the same stimulus words as the HA paragraph, but the words designated as response words in the HA paragraph were deleted. In their place, for the LA paragraph, words were used which are not commonly associated with the stimulus words. These words were of the same Thorndike-Lorge frequency, were semantically appropriate in the context, and were grammatically the same as the words they replaced.

To test recall, 12 multiple-choice questions were written. The same questions were given to all *S*s regardless of the paragraph they read. Each question had four alternatives. One of the four alternatives was a response word from the HA paragraph while a second alternative was a response word from the LA paragraph. The two other alternatives were used as distractors. To answer the question correctly, *S* had to select the alternative containing the response word from the paragraph he read.

Parts of the HA and LA paragraphs are reproduced below along with one of the questions.

HA. They were all happy to be together again. Outside the moon and stars shone brightly in the June sky, and the green grass sparkled in the night.

LA. They were all relieved to be together again. Outside the moon and lake appeared clearly in the June evening, and the green house sparkled in the valley.

Question. The green _____ sparkled. (a. house, b. plants, c. grass, d. emeralds)

The paragraphs and questions were mimeographed on 8½ × 11-inch paper. Page 1 contained either the HA or LA paragraph, while Pages 2 and 3 contained the questions.

Procedure. The experimenter (*E*) worked individually with *S*s. The *S* was told that he was going to be given a paragraph to read, and that after reading the paragraph he would have to answer questions about the paragraph. The *S* was also told to read the paragraph quickly but carefully, and that *E* would not be able to help *S* read any words. The *S* was told to look up to indicate he was finished reading. At a signal from *E*, *S* began to read. A stopwatch was used to measure the time required to read the paragraph. To answer questions, *S* was instructed to circle the alternative he thought was correct.

The *S*s in the control condition were tested in a group. Their materials contained only the 12 questions. They were told: "Answer these questions as you think they should be answered."

Results

The mean time to read the HA paragraph was 43.82 seconds ($SD = 6.62$) while the mean time to read the LA paragraph was 58.81 seconds ($SD = 36.73$). Although the t test is a robust test with regard to moderate departures from assumptions regarding homogeneity of the variances, the variances in this analysis were sufficiently different to warrant the use of Welch's (Winer, 1962) approximation to the sampling distribution of the t statistic. Using the conservative test, the difference between the means was significant ($t = 2.04$, $df = 27$, $p < .05$, one-tailed).

The mean number of questions answered correctly for the HA group was 9.50 ($SD = 1.43$), while the mean number of questions answered correctly by the LA group was 5.04 ($SD = 1.91$). The difference between the means of the two groups was significant ($t = 9.70$, $df = 52$, $p < .001$, one-tailed).

When the mean number of HA alternatives correctly chosen by HA *S*s who read the paragraph ($M = 9.50$, $SD = 1.43$) was compared with the mean number of HA alternatives guessed by control *S*s ($M = 5.75$, $SD = 1.35$), the difference was significant ($t = 11.03$, $df = 66$, $p < .001$, one-tailed).

When the mean number of LA alternatives correctly chosen by LA *S*s who read the paragraph ($M = 5.04$, $SD = 1.91$) was compared with the mean number of LA alternatives guessed by control *S*s ($M = .88$, $SD = .85$), the difference was significant ($t = 10.49$, $df = 64$, $p < .001$, one-tailed).

To test the prediction that when forced to guess, *S*s would choose alternatives on the basis of the associative strength of the alternatives, a frequency distribution of choices for high- and low-association strength alternatives was made for each question. The high-association strength alternative was chosen more frequently for 11 of the 12 questions (sign test, $p < .001$, one-tailed).

EXPERIMENT II

Method

Subjects. The *S*s were 135 juniors enrolled in an educational psychology course. Fifty *S*s were in the HA, 53 *S*s in the LA, and 32 *S*s in the control condition.

Materials. The same materials were used for the three conditions as in Experiment I.

Procedures. HA and LA *S*s were tested at the same time. After the *S*s were seated, they were told not to open their test materials until instructed. The test materials were then randomly distributed by alternately placing HA then LA materials face down on the desks of *S*s. The *S*s were then told that a signal would be given upon which they were to begin to read the paragraph quickly and carefully. Before they began to read, they were told that as soon as they finished reading the paragraph, they were to find the number of seconds it had taken them to read the paragraph by looking at the chalk board located in the front of the room. Upon the board consecutive numbers were written. The *S*s found the time it had taken them to read the paragraph by locating the last number crossed off the board, and they wrote this number at the top of Page 2 of the test materials. The *E* held a stopwatch, and as each second passed he crossed off the number indicating the time. The *S*s were told that once they finished reading the paragraph they could not go back to the paragraph. Answers to the questions were to be indicated by circling one of the alternatives.

The *S*s in the control condition were tested as a group at a different time. Instructions were the same as those given to control *S*s in Experiment I.

Results

The mean time to read the HA paragraph was 35.26 seconds ($SD = 7.66$) while the mean time to read the LA paragraph was 38.86 seconds ($SD = 6.45$). The difference between the means of the two groups was significant ($t = 2.57$, $df = 101$, $p < .01$, one-tailed).

The mean number of questions answered correctly by the HA group was 9.68 ($SD = 1.81$), while the mean number of questions answered correctly by the LA group was 6.87 ($SD = 2.30$). This difference between the means of the two groups was significant ($t = 6.89$, $df = 101$, $p < .001$, one-tailed).

When the mean number of HA alternatives correctly chosen by the HA *S*s who read the paragraph ($M = 9.68$, $SD = 1.81$) was compared with the mean number of HA alternatives guessed by control *S*s ($M = 6.03$, $SD = 4.29$), the difference was significant ($t = 8.11$, $df = 80$, $p < .001$, one-tailed).

When the mean number of LA alternatives correctly chosen by LA *S*s who read the paragraph ($M = 6.87$, $SD = 2.30$) was compared with the mean number of LA alternatives guessed by control *S*s ($M = 1.59$, $SD = 1.16$), the difference was significant ($t = 14.64$, $df = 83$, $p < .001$, one-tailed).

To test the prediction that when forced to guess, *S*s would choose alternatives on the basis of the associative strength of the alternatives, a frequency distribution of choices for high- and low-association strength alternatives was made for each question. The high-association strength alternative was chosen more frequently for 10 of the 12 questions (sign test, $p < .003$, one-tailed).

DISCUSSION

The results supported the hypotheses that word-association strength influences reading speed, recall, and guessing behavior on multiple-choice tests. The paragraph with stronger associative relationships between words was read faster and with better recall than the paragraph with weaker associative relationships. When forced to guess on multiple-choice questions, *S*s tended to choose alternatives that had strong associative relations with words in the stem of the question.

Although these results support the hypotheses, the mechanisms through which word associations affect reading are not disclosed. Eye movement photography taken while reading indicates that eye movements consist of fixations, interfixations, and regressions. Fixations account for about 92–94% of the reading time. Gilbert (1959) separated duration of fixation into seeing time, central processing time, and stabilizing time. The length of time given to a fixation depends on the difficulty of the passage (Tinker, 1958) and familiarity with the textual material (Morton, 1964). A regression is a return to a previously fixated word and occurs when there is need for verification (Bayle, 1942). Whatever effect word associations may have on the mechanisms just described, an assumption is made that word associations have little effect on stabilizing and interfixation time, but may influence reading speed by affecting seeing time, central processing time, and number of regressions.

REFERENCES

BAYLE, E. The nature and causes of regressive movements in reading. *Journal of Experimental Education,* 1942, **11,** 16–36.

DEESE, J. Influence of inter-item associative strength upon immediate free recall. *Psychological Reports,* 1959, **5,** 305–312.

GILBERT, L. C. Speed of processing visual stimuli and its relation to reading. *Journal of Educational Psychology,* 1959, **50,** 8–14.

GOODMAN, K. S. A linguistic study of cues and miscues in reading. *Elementary English,* 1965, **42,** 639–643.

MORTON, J. The effect of context upon speed of reading, eye movements and eye-voice span. *Quarterly Journal of Experimental Psychology,* 1964, **16,** 340–354.

O'NEIL, W. M. The effect of verbal association on tachistoscopic recognition. *Australian Journal of Psychology,* 1953, **49,** 333–338.

PALERMO, D. S., & JENKINS, J. J. *Word association norms—grade school through college.* Minneapolis: University of Minnesota Press, 1964.

ROSENBERG, S. Associative factors in the recall of connected discourse. In, progress report for *Linguistic structure as a variable in verbal learning,* 1965. Nashville: George Peabody College for Teachers, Department of Psychology. (mimeo.)

ROUSE, R. O., & VERNIS, S. J. The effect of associative connections on the recognition of flashed words. *Journal of Verbal Learning and Verbal Behavior,* 1963, **1,** 300–303.

TAYLOR, S. E. Sensation and perception: The complexity of word perception. *Journal of Developmental Reading,* 1963, **6,** 187–206.

TINKER, M. A. Recent studies of eye movements in reading. *Psychological Bulletin,* 1958, **55,** 215–231.

TULVING, E., & GOLD, C. Stimulus information and contextual information as determinants of tachistoscopic recognition of words. *Journal of Experimental Psychology,* 1963, **66,** 319–327.

WINER, B. J. *Statistical principles in experimental design.* New York: McGraw-Hill, 1962.

(Received January 23, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 1, 16-19

EFFECTS OF REVIEW AND TESTLIKE EVENTS WITHIN THE LEARNING OF PROSE MATERIALS

ROGER H. BRUNING[1]
University of Nebraska

Effects of format and criterion-test (CT) relevance of within-learning review items were tested in a factorial design. A 1500-word passage was divided into 6 sections, each followed by from 3 to 7 statement- or test-type review items. Test-type and CT-relevant groups showed significantly fewer errors ($p < .01$) than statement-type and CT-irrelevant groups, respectively. CT-irrelevant testing reduced errors ($p < .01$) over both CT-irrelevant statements and a reading control group, indicating that testing independently facilitates learning from such materials. Additionally, effects of testing and specific review were seen to be additive with CT-relevant testing significantly reducing errors ($p < .01$) over both CT-relevant statement and CT-irrelevant tested review.

Recent research (Hershberger & Terry, 1965a, 1965b; Rothkopf, 1966) has indicated that testlike events within prose learning materials may have both specific and general facilitative effects on learning. Specifically, such events may function as additional practice trials for factual and conceptual materials and they may also serve more generally as one type of environmental control for inspection behaviors, affecting such things as set, rate, and persistence of the reading responses. Under the control of such testing the learner tends to engage in more careful inspection of the prose document and to search for meaningful facts and concepts consistent with those encountered in the testlike events.

An earlier study by Rothkopf and Coke (1963) presents evidence that such test-type review may have a relatively greater positive effect on learning than does review of identical materials in the form of declarative sentences. However, results were positive only for a mixed-mode design in which subjects (*S*s) learned materials under both review conditions, and when independent treatment groups were employed, sentence review was at least as good as test-type review. The present study represents an additional test of the relative efficacy of sentence-type and test-type review when independent treatment groups are employed.

Although Rothkopf's (1966) study presents strong indication that such testlike events function both independently and in conjunction with review content, no measures of the relative and combinational effects of the two variables are available since different criterion measures were used for each. To provide additional evidence on this question, the present study employed a single test of retention to permit a more direct comparison of the effects of testlike events per se with those of testlike events which serve the additional function of reviewing materials relevant to the criterion test. A 2 × 2 factorial analysis of variance design was thus employed to test the effects of the manipulated parameters, review relevance and review item format.

METHOD

Subjects. The *S*s were 69 students from an introductory educational psychology course at the University of Nebraska. Participation in the experimental sessions was a course requirement.

Materials. A highly factual, descriptive passage of approximately 1500 words dealing with characteristics of a fictitious African tribe was formulated. The passage was divided into six parts, the first approximately 150 words long and the others approximately 275 words each. Following each section was a review page containing from three to seven items dealing with facts presented in that section.

Procedure. Variations in the material and for-

[1] The author would like to express his appreciation to Kenneth D. Orton, University of Nebraska, for his most helpful suggestions throughout the study and for his critical reading of the manuscript.

mat of the review pages constituted the major experimental treatments of the study. From six to fourteen statements of major factual points were drawn from each section. The statements from each section were then paired by the experimenter on criteria of difficulty and importance. Members of pairs were then assigned at random to either a general criterion test-relevant (CR) or to a criterion test-irrelevant (CI) condition. For each relevance condition two formats for review items were developed, statement-type items and test-type items. Statement-type items consisted of declarative sentences while test-type items on the same relevance dimension were identical except for the deletion of the key response term, which *S* was required to supply. Knowledge of results for test-type items was provided by lifting a tab covering the answer. As combinations of the major variables, criterion relevance and format, the four major treatment conditions were: (*a*) criterion-irrelevant statements (CIS), (*b*) criterion-irrelevant test items (CIT), (*c*) criterion-relevant statements (CRS), and (*d*) criterion-relevant test items (CRT). A fifth reading control group ($N = 13$) had no review material of any kind within the prose materials. This group only read through the materials sequentially without review or testing within the learning situation. For all *S*s, materials were learned under only a single treatment condition.

The general criterion test (CT) was composed of the 26 review items associated with the final five sections of the learning material and previously assigned to the CR conditions. Thus these items were identical in content to the statements of the CRS condition and completely identical in both content and format to the test items of the CRT condition. On the CT no knowledge of results was provided, however, and items were randomized to control for possible serial effects associated with the original learning materials.

Learning and review materials were put into booklet form and coded by treatment condition. Booklets were then randomized and handed out in a classroom setting. General instructions required that *S*s read through the materials at their own speed in preparation for a posttest and, in addition, did not allow *S*s to skip ahead or to look back once they had finished with a page. Instructions specific to learning conditions were contained in the booklets. For CIS and CRS (statement format) conditions review pages were prefaced by the following instruction:

> As a review, some of the major points made on the preceding page are the following.

For CIT and CRT (test format) conditions, review pages were prefaced by the following:

> Try to answer these questions. After writing the answer in the space, check it by lifting the tab across from the answer. Please do not look at the answer until you have answered on your own. These questions are only for review and will not be graded. Answers may require more than one word in some cases.

When an *S* had finished with the learning materials, his booklet was taken up, time taken to complete the materials was noted, and a copy of the CT given him. The CT was prefaced by a short questionnaire requiring some personal data and a rating of the difficulty of the reading materials and, as result, a period of approximately 5 minutes elapsed between the completion of learning and beginning on the CT itself. The major purpose of the questionnaire was to interpolate a short period of time between learning and tested recall, thus minimizing the possibility that answers could be drawn from the immediate memory store.

RESULTS

Mean error scores of the five treatment conditions are presented in Table 1. As can be seen from the range of scores, the performance on the criterion measure varied considerably as a result of the different treatment conditions. In the analysis of variance for the four major treatment conditions, main effects for both criterion relevance and format conditions were significant beyond the .01 level ($F = 31.18$, $df = 1/52$, and $F = 17.89$, $df = 1/52$, respectively), although there were virtually no interaction effects. Contrasts of the individual means (Scheffé, 1959) on the test format level revealed significantly fewer errors ($p < .01$) for the CRT than for the CIT condition and with the statement format, the CRS group showed significantly fewer errors ($p < .01$) than the CIS group. Similarly, in comparisons by level across the format dimension, CRT scores were significantly lower ($p < .01$) than CRS scores and CIT scores significantly lower than CIS scores ($p < .01$). Thus these results substantiate those of Rothkopf (1966) in showing that such testing within learn-

TABLE 1

MEAN ERROR SCORES ON THE CT AND MEAN TIME (IN MINUTES) SPENT ON THE LEARNING MATERIALS

Treatment	Errors		Time	
	Mean	*SD*	Mean	*SD*
CIS	14.93	6.18	14.50	2.68
CRS	9.07	2.86	14.42	2.67
CIT	10.50	3.01	20.79	3.91
CRT	4.57	2.62	21.93	3.61
RC	14.00	3.85	11.84	1.46

ing, even when content tested is unrelated to that appearing on the criterion measure, independently and significantly improves performance on a criterion test. That the effects of review and testing may be additive is indicated in the difference between the CRT condition and both the CRS and CIT conditions. While both criterion-relevant review by statements and criterion-irrelevant testing brought about substantial gains in learning over the CIS condition, their combined effects were significantly greater than either alone.

A potential source of bias in the analysis was an inequality of cell variances. However, results from Scheffé (1959) indicate that inequalities of the order existing in the present data with equal cell sizes cause only a slight increase in the probability of a Type I error and would not appear to nullify the present results since all comparisons were significant beyond the .01 level. As an additional check, however, logarithmic transformations of scores were employed in a separate analysis. Results from these scores completely substantiated those using raw score data with effects of format and criterion relevance both significant beyond the .01 level (F = 18.6, 29.2; df = 1/52), together with a similar nonsignificant interaction effect.

The t-test comparisons of the reading control group (RC) with the four experimental conditions at the .01 level revealed that all treatment conditions except the CIS condition had significantly lower error rates on the CT than did the RC group.

As a comparison of the relative potency of the two major treatment variables, criterion relevance and format, an estimation of the correlation between the variables and criterion measure, omega squared (ω^2), (Hays, 1963) was obtained. For the relevance dimension, ω^2 was .30 compared with .17 for the format dimension. Thus, in terms of explained variance, the relevance of review can be seen as contributing a somewhat larger proportion of explained criterion variation than did the format manipulations. However, as is evident from the results of Table 1, mean differences between the conditions of testing without review (CIT) and the condition of review without testing (CRS) were slight and, of course, nonsignificant.

An analysis of time scores for the experimental conditions revealed a significantly greater (F = 60.3, df = 1/52, $p < .001$) amount of time spent on the learning and review materials for test format conditions and a predicted nonsignificant effect on the relevance dimension. The increased time spent on the testing format conditions was due in part, of course, to the time spent in formulating and writing out responses, but the significant reduction of errors associated with the testing format points to an increased amount of time spent in the inspection of the prose document (Rothkopf, 1966).

DISCUSSION

The strong facilitating effect of adjunct testing upon learning found in the present study gives additional support to Rothkopf's (1965) hypothesis that such infrequent testing within the learning of prose materials may be an important environmental control of such learning behaviors. Also, the present results show that such testing has facilitative effects both in conjunction with and independent of the review of the actual content to be learned. In highly factual material of the type employed in the present study, interrelationships among content areas were minimal and transfer effects from material tested within learning on the CI dimension and the material appearing on the posttest would seem correspondingly very small. That such specific transfer effects from the "irrelevant" review to the criterion test were indeed inconsequential is indicated in the performance of the CIS group, which actually showed a slightly higher error rate than did the reading control group, which had no review of any kind. Thus the positive effects found when review was irrelevant and test format was employed can be attributed only to improvement in learning behaviors associated with the prose materials caused by the testing itself; increased per-page inspection time, more active reading behavior, self-prompting and the like, together with any generalized reinforcing effects which may

have occurred due to self-testing. When review was directly relevant to the criterion test, the significant reduction of errors through testing would seem due in part to these same general factors. In addition, however, the facilitating effects of active formulation, rehearsal, and reinforcement of specific responses would appear to be operant.

As has been noted previously (Rothkopf, 1965), the principles of active formulation and testing of responses within learning have been commonly employed within the context of programmed instruction. However, a very high test/content ratio exists in programmed instruction together with rigid behavioral controls in contrast to the very low test/content ratio and the more informal controls exerted in the present prose materials. Nevertheless, in spite of the apparently minimal controls which are applied, such infrequent testing is seen to bring about significant increases in the learning from prose documents, at least within the experimental context. If further research reveals a wider generality in its functioning to various grammatical and contextual situations, such testing may present a possible and perhaps highly practical alternative to the lack of effective behavioral control in ordinary reading and to the inherent rigidity of programmed instruction.

REFERENCES

HAYS, W. L. *Statistics for psychologists.* New York: Holt, Rinehart & Winston, 1963.

HERSHBERGER, W. A., & TERRY, D. F. Delay of self-testing in three types of programed texts. *Journal of Educational Psychology,* 1965, **56,** 22–30. (a)

HERSHBERGER, W. A., & TERRY, D. F. Typographical cueing in conventional and programed texts. *Journal of Applied Psychology,* 1965, **49,** 55–60. (b)

ROTHKOPF, E. Z. Some theoretical and experimental approaches to problems in written instruction. In J. D. Krumboltz (Ed.), *Learning and the educational process.* Chicago: Rand McNally, 1965. Pp. 193–221.

ROTHKOPF, E. Z. Learning from written instructive materials: An exploration of the control of inspection behavior by test-like events. *American Educational Research Journal,* 1966, **3,** 241–249.

ROTHKOPF, E. Z., & COKE, E. U. Repetition interval and rehearsal method in learning equivalences from written sentences. *Journal of Verbal Learning and Verbal Behavior,* 1963, **2,** 406–416.

SCHEFFÉ, H. *The analysis of variance.* New York: Wiley, 1959.

(Received January 26, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 1, 20–25

TEXTUAL CONSTRAINT AS FUNCTION OF REPEATED INSPECTION

ERNST Z. ROTHKOPF[1]
Bell Telephone Laboratories, Incorporated
Murray Hill, New Jersey

College and high school students (N = 155) were exposed to 1 of 2 experimental passages either 0, 1, 2, or 4 times. After the completion of the required number of self-paced inspections, learning (textual constraint) was measured by the completion method (Cloze procedure). Repeated exposure was accompanied by declines in both practice text and completion test inspection times. Proportion of correct fill-in responses was found to be an increasing, negatively accelerated function of the number of practice exposures. The data were consistent with the view that repeated, massed exposures resulted in progressive modification or extinction of inspection (mathemagenic) behavior. 2 alternative hypotheses were rejected. The completion procedure appears to be a simple, quantitative method for estimating what is learned from written discourse.

Textual constraint has been estimated by guessing procedures such as those proposed by Shannon (1951), and others (e.g., Burton & Licklider, 1955; Miller & Friedman, 1957). With connected prose the guessing units have usually been single words which have been deleted from the text and which subjects (*S*s) are asked to replace. This method, which has sometimes been called the completion procedure, has been used in psychological investigations at least since 1897 (Ebbinghaus, 1897; for a historical discussion see MacGinitie, 1960). More recently, it has become increasingly common to follow the practice of Taylor (1954) and refer to this method as the Cloze procedure.

Studies of constraints in language have been motivated in the past mainly by questions about information processing and similar short-term effects of communications. Some investigators, for example, have been trying to specify how much noise a language-encoded message can tolerate before it becomes difficult to understand (e.g., Miller, Heise, & Lichten, 1951). Others have been trying to predict the *readability* of books (Taylor, 1953) or to understand the information-handling capacity of readers (Pierce & Karlin, 1957).

Implicit in the notion of textual constraint is the assumption that these constraints reflect learned associations among various units of the language chain. These learned associations can be thought to belong to two general classes: (*a*) those related to syntax, which are relatively similar from passage to passage within a given language; and (*b*) those related to the guesser's knowledge about the subject matter with which the experimental passage is concerned. The second class of constraining associations can be expected to vary widely from topic to topic, passage to passage, and from reader to reader.

The hypothesized character of the second source of textual constraint suggests that the completion method may provide a suitable measure of learning from connected discourse. The usefulness of the completion method in measuring the acquisition of complex verbal skills is of considerable interest because of the need for simpler quantitative performance measures in research on instructional methods and on complex learning.

Systematic studies of the simple effect of practice exposures on textual constraint, as measured by the completion (Cloze) method, do not appear to be available in the experimental literature. Carroll and associates (Carroll, Carton, & Wild, 1959) did demonstrate that textual constraint in Russian text increased after training in Russian, but amount of practice was not systematically varied in their study. King

[1] The author is greatly indebted to Esther U. Coke for help in the computer analysis of the data. Carolyn Shefsky provided valuable assistance in all phases of this experiment.

and Cofer (1960) aurally presented two short stories six times and found increases in Cloze performance that leveled off at about 60% correct responses after six presentations. However, the effects of exposure to the experimental passage and repeated completion testing were completely confounded in their experiment. As a consequence it was not possible to determine the simple effect of text exposure.

The purpose of the present experiment was to determine changes in textual constraint that result from practice exposures when the confounding effects of intermittent testing have been eliminated. The completion method was again used to provide a measure of constraint. The study was carried out under conditions in which exposure time on any one presentation of the experimental text was under the control of *S*. This was done because people usually control their own exposure time during the study of written text. Therefore frequency of exposure with self-pacing appeared a more interesting and useful independent (practice) variable than exposure frequency with time per exposure held constant. The use of self-paced practice was also partly motivated by interest in inspection time as a dependent measure.

METHOD

Materials

Two passages, one on leather making (approximately 1500 words, excerpted from Parker, 1911,[2]) and one on the history of Australia (approximately 750 words, excerpted from MacInnes et al., 1964, pp. 23–24[3]) were used. Practice and completion test material were mimeographed. Since "substantive" or "content" learning was the focus of interest, only nonfunction words were deleted in the completion test. With this restriction, approximately one word in ten was removed by an otherwise unbiased algorithm. A line of uniform length was put in place of the deleted word.

Design and Procedure

Each *S* was exposed to only one of the two experimental passages, either 0, 1, 2, or 4 times. For the leather passage, numbers of *S*s for the four exposure frequencies were respectively 16, 17, 18, and 17. For the text on Australia these numbers were 22, 24, 21, and 20. Self-pacing and group procedures were used throughout. Whenever *S* completed an inspection of the experimental passage, he signaled the experimenter, who then collected the text and immediately provided the document which was required for the next phase of the experiment.

Approximately 10 minutes after the completion of the predetermined number of inspections of the passage, *S*s were given the completion test. During the 10-minute delay, *S*s completed a personal questionnaire.

The *S*s were instructed to record the time they started and finished each page of text. They also recorded these times for the completion test. Time was recorded from a digital clock which was projected on a screen.

Subjects

All *S*s who read the leather passage were paid volunteer undergraduate students from Fairleigh Dickinson University in Madison, New Jersey. So were half of the *S*s reading the text on Australia. The remaining Australia *S*s were paid volunteers from Grades 11 and 12 of Cranford High School, Cranford, New Jersey.[4]

RESULTS AND DISCUSSION

Correct Completion Responses

Proportion of correct responses over all items on the completion test is shown for the various treatments in Figure 1. A re-

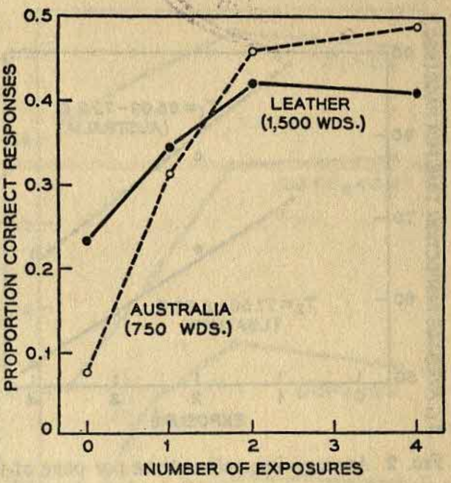


FIG. 1. Proportion of correct responses on the completion test as function of number of exposures to the experimental passage.

[2] This excerpt was used with the kind permission of the Encyclopaedia Brittanica, 425 North Michigan Avenue, Chicago, Illinois.

[3] This material was used with the kind permission of Time, Inc., Rockefeller Center, New York, New York.

[4] The author would like to thank Charles Post, the vice principal of the Cranford High School, for his help in obtaining these *S*s.

Acc. 5679
Date 18.3.6
(S. C. E. R. T.)

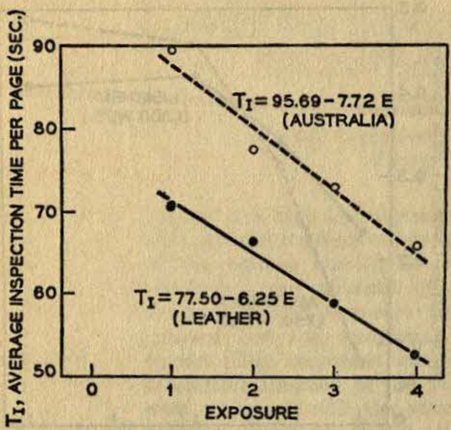


FIG. 2. Average inspection time per page of text for the several exposures. (Only data from the 4-exposure groups are shown. The lines were fitted by the method of least squares, E = exposure.)

sponse was scored correct if it was identical to the word which had been deleted from the text. Minor variations in spelling were accepted as correct. Correct responses increased as a function of repeated exposure to the passage but the learning curve was negatively accelerated and leveled off after two inspections. The plots in Figure 1 are quite similar to the acquisition curves which have been obtained from averaged group data for many other learning tasks.

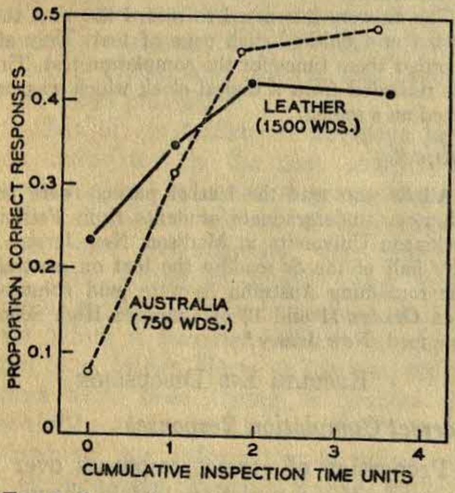


FIG. 3. Proportion of correct responses on the completion test as function of the cumulative time spent in inspecting the passage. (Time is shown in multiples of inspection time on Exposure 1—see text for explanation.)

The rise in textual constraint with repeated exposure was tested by analysis of variance. The results were significant for both the Leather ($F = 26.53$; $df = 3/64$; $p < .001$) and the Australia passages ($F = 39.07$; $df = 3/83$; $p < .001$).

Inspection Time

One possible account for the negatively accelerated form of the acquisition curves in Figure 1 is that it is related to the decreases in inspection time which occurred at each successive exposure. Decline of inspection time with repetition or prolonged reading have been observed previously (Premack & Collier, 1966; Rothkopf & Bisbicos, 1967). Average inspection time per page is shown in Figure 2 for *S*s who were exposed to the experimental passage four times. The drop in inspection time was approximately linear and, as tested by analysis of variance, with repeated observations on the same *S*s, was significant for both the Leather ($F = 8.73$; $df = 3/45$; $p < .01$) and the Australia ($F = 3.32$; $df = 3/36$; $p < .05$) passage. Only those *S*s for whom inspection times were available for all pages over all four readings were used for these analyses.

However, the gain from successive inspections declines at a greater rate than inspection time. This is illustrated in Figure 3, where proportion of correct responses on the completion test was plotted as a function of cumulative inspection time through four exposures. Because of differences among treatment groups in average inspection time on the first exposure to the passage, cumulative inspection time was plotted as a multiple of the amount of time spent on the first reading. It is quite clear from inspection of Figure 3 that less and less learned constraint resulted from successive inspection time units as *S*s were repeatedly exposed to the experimental passages.

The results shown in Figure 3 indicated that the amount of learning produced by each unit of inspection time decreases throughout the several inspections. The observed decline in inspection time over exposures therefore cannot provide a suffi-

cient account for the leveling off of the acquisition curves in Figure 1. Another explanation may be that the amount learned at each successive inspection is proportional to the amount which has yet to be learned. Conceptual models which employ the proportionality notion have been used to account for the negatively accelerated acquisition curves of several learning tasks (e.g., Estes, 1950; Hull, 1943, pp. 102–123). One of the consequences of such a model is that learning curves for items of various levels of difficulty should differ markedly from each other in slope. This prediction was tested, for the Leather passage,[5] by subdividing the 123 items of the Leather completion test according to p_0, the initial guessing difficulty, that is, percentage correct response for *S*s who did not see the passage prior to the completion test. Four different levels were used: $p_0 = 0(N = 45)$, $0 < p_0 < 0.2(N = 29)$, $0.2 < p_0 < 0.4(N = 18)$, $p_0 > 0.4(N = 31)$. Proportion of correct completion responses as a function of number of inspections prior to testing is shown for these four categories of items in Figure 4. The four curves were nearly parallel. Each of the four classes of items gained, on the average, about the same number of correct responses per exposure.

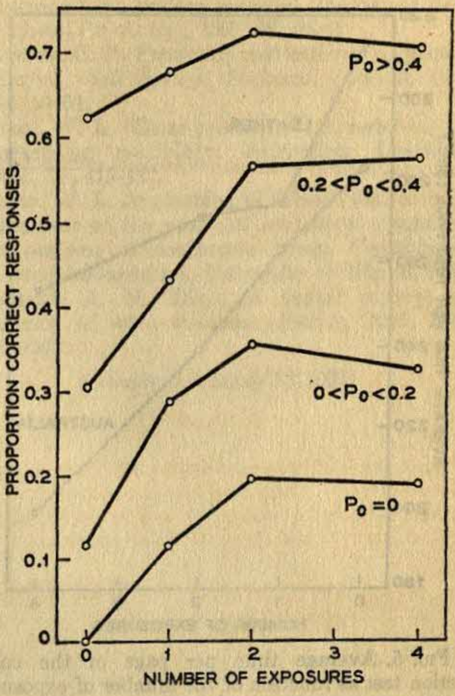


FIG. 4. Proportion of correct responses on the completion test as function of number of exposures to the experimental passage for test items of various 0-exposure guessing probabilities (P_0). (The data are from the Leather passage.)

The present analysis therefore *did not* support interpretation of slowing of the learning rate in terms of approach to a common theoretical maximum constraint for all items such as "mastery." The additional assumption that theoretical maxima differ for various deletions in the passage, however, may serve to bring the proportionality hypothesis in line with the data. This is not an unreasonable assumption since it is well known that there are large differences in ambiguity among various syntactic and semantic constructions (e.g., Aborn, Rubinstein, & Sterling, 1959; Coleman, 1965).

A more likely explanation may be based on the modification or extinction of attention-like processes as a function of repeated or prolonged inspection (see Rothkopf, 1966). Learning from written materials depends critically on a class of attention-like responses, called mathemagenic behaviors or inspection behaviors (Rothkopf, 1963, 1965). The extinction or modification of these attention-like processes by repeated, massed inspection would result in the diminishing learning effects. The decline in inspection time, according to this interpretation, is an indicator of the successive changes in mathemagenic behavior. More direct tests of this hypothesis are now underway.

Completion Test Time

Average test time per page of the completion test was plotted as a function of number of inspections of the text (Figure 5). Test time generally decreases for both passages. The data are well fitted by the line $T = 283.57 - 21.04\,E$ for Australia, where T is test time and E is exposure.

[5] The Leather passage was used exclusively because the number and distribution of data points for the Australia passage was not sufficient.

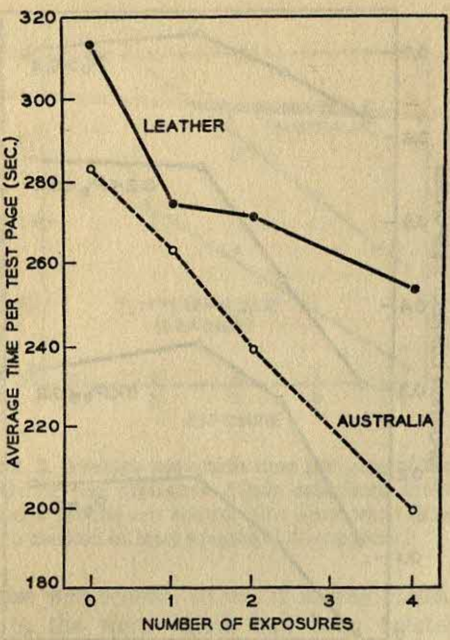


FIG. 5. Average time per page of the completion test as function of the number of exposures to the experimental passage.

The basis for the decline in test time cannot be decided from the present results. It may be due to (*a*) decreased latencies in filling in the blanks, or (*b*) decreased time in inspecting the remaining reading materials in the test, or both. The first effect, that is, a decrease in response latency, has been frequently observed to be associated with rises in the probability of correctly learned responses (e.g., Hull, 1943, pp. 336–337). It has also been specifically observed in connection with increases in verbal context (Treisman, 1965). The second possibility is that test times decrease as function of exposure to the practice passage because the reading material in the test is inspected more hurriedly. This suggests that the decreases in inspection time of the practice and the test material may have a common basis.

Both of the factors described above may contribute to the decline in the time it takes to complete the completion test. The present data do not rule out the conclusion that repeated exposure to an experimental passage results not only in more hurried inspection of the practice passage but also of the test material. The possibility therefore exists that the negative acceleration of the constraint learning curve was at least in part due to changes in test inspection behaviors rather than modification of mathemagenic behavior during practice.

REFERENCES

ABORN, M., RUBINSTEIN, H., & STERLING, T. D. Sources of contextual constraint upon words in sentences. *Journal of Experimental Psychology,* 1959, **57,** 171–180.

BURTON, N. G., & LICKLIDER, J. C. R. Long-range constraints in the statistical structure of printed English. *American Journal of Psychology,* 1955, **68,** 650–653.

CARROLL, J. B., CARTON, A. S., & WILD, C. P. An investigation of "cloze" items in the measurement of achievement in foreign languages. Cambridge: Laboratory for Research in Instruction, Graduate School of Education, Harvard University, 1959.

COLEMAN, E. On understanding prose: Some determiners of its complexity. Unpublished report, New Mexico State University, 1965.

EBBINGHAUS, H. Über eine neue Methode zur Prüfung geistiger Fähigkeiten und ihre Anwendung bei Schulkindern. *Zeitschrift für Psychologie,* 1897, **13,** 407–424.

ESTES, W. K. Towards a statistical theory of learning. *Psychological Review,* 1950, **57,** 94–107.

HULL, C. L. *Principles of behavior.* New York: Appleton-Century-Crofts, 1943.

KING, D. On the accuracy of written recall; a scaling and factor analytic study. *Psychological Record,* 1960, **10,** 113–122.

KING, D., & COFER, C. N. Exploratory studies of stories varying in the adjective-verb quotient. *Journal of General Psychology,* 1960, **62,** 199–221.

MACGINITIE, W. H. Contextual constraint in English prose. Unpublished doctoral dissertation, Columbia University, 1960.

MACINNES, C., et al. *Australia and New Zealand,* Life World Library, New York: 1964.

MILLER, G., & FRIEDMAN, E. A. The reconstruction of mutilated English texts. *Information and Control,* 1957, **1,** 38–55.

MILLER, G. A., HEISE, G. A., & LICHTEN, W. The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology,* 1951, **41,** 329–335.

PARKER, J. G. Leather. In *The encyclopaedia Britannica,* (11th ed.). Cambridge, England: Cambridge University Press, 1911. Vol. XVI. Pp. 330–345.

PIERCE, J. R., & KARLIN, J. E. Reading rates and the information rate of a human channel. *Bell System Technical Journal,* 1957, **36,** 497–516.

PREMACK, D., & COLLIER, G. Duration of looking and number of brief looks as dependent variables. *Psychonomic Science,* 1966, **4,** 81–82.

ROTHKOPF, E. Z. Some conjectures about inspection behavior in learning from written sentences and the response mode problem in programed self-instruction. *Journal of Programed Instruction,* 1963, **2,** 31–46.

ROTHKOPF, E. Z. Some theoretical and experimental approaches to problems in written instruction. In Krumboltz, J. D. (Ed.), *Learning and the educational process.* Chicago: Rand McNally, 1965. Pp. 193–221.

ROTHKOPF, E. Z. Learning from written material: An exploration of the control of inspection behavior by test-like events. *American Educational Research Journal,* 1966, **3,** 241–249.

ROTHKOPF, E. Z., & BISBICOS, E. E. Selective facilitative effect of interspersed questions on learning from written material. *Journal of Educational Psychology,* 1967, **58,** 56–61.

SHANNON, C. E. Prediction and entropy of printed English. *Bell System Technical Journal,* 1951, **30,** 50–64.

TAYLOR, W. L. "Cloze procedure"; A new tool for measuring readibility. *Journalism Quarterly,* 1953, **30,** 415–433.

TAYLOR, W. L. Application of "cloze" and entropy measures to the study of contextual constraint in samples of continuous prose. Unpublished doctoral dissertation, University of Illinois, 1954.

TREISMAN, A. M. Effect of verbal context on latency of word selection. *Nature,* 1965, **206,** 218–219.

(Received January 30, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 1, 26–31

GRADE LEVEL, SCHOOL STRATA, AND LEARNING EFFICIENCY[1]

WILLIAM D. ROHWER, JR., STEVE LYNCH, JOEL R. LEVIN, AND NANCY SUZUKI[2]
University of California, Berkeley

A 4-way design was used to evaluate the facilitory effects of sentence verbalization and action depiction on the learning of paired associates by 1st-, 3rd-, and 6th-grade children from high- and low-strata schools. Each of a total of 432 Ss learned a list of 24 pictures of paired objects presented on movie film by a pairing-test procedure for 2 trials. The 1st of 2 experimental variables, Verbalization, concerned the type of verbal description given for the pairs (names, phrases, sentences). The 2nd experimental variable, Depiction, contrasted 2 kinds of pictorial representations of the pairs, one of which was a visual translation of the name and phrase descriptions (still) and the other of which was a visual translation of the sentence descriptions (action). The amount learned by older Ss was greater than that learned by younger Ss regardless of school strata. Sentence descriptions and action depiction facilitated learning for all Ss, and, in all conditions, low-strata children learned as efficiently as high-strata children.

The present experiment was performed in order to evaluate hypotheses suggested by the juxtaposition of two rather disparate topics of current research interest: The improvement of learning efficiency, and group-related differences in learning efficiency. Concern with the former topic is well illustrated in the work of Davidson (1964), Jensen and Rohwer (1965), and Reese (1965) on conditions for the facilitation of paired-associate (PA) learning in children. Two kinds of facilitory conditions have been isolated: verbal and pictorial. Jensen and Rohwer found that the acquisition of a list of paired pictures by second-, fourth-, and sixth-grade children was markedly accelerated by the instruction to form and utter a sentence containing the names of the two objects in each pair. In this and a number of subsequent experiments (Rohwer, 1966; Rohwer & Lynch, 1966; Rohwer, Shuell, & Levin, 1968) the facilitory effect of sentence contexts has been clearly demonstrated. On the pictorial side, Davidson (1965) has shown that the learning of paired objects represented by line drawings is determined by the spatial configuration of the two members in a pair. When the two objects in each pair were depicted independently, the amount learned was notably smaller than when the two objects were joined to one another visually (e.g., a picture of a chain and a bowl, versus a picture of a chain in a bowl). Reese (1965) found that verbal descriptions of relationships between objects as well as pictorial representations of such relationships facilitated PA learning in young children.

The samples of children involved in all of the experiments reviewed thus far were drawn from schools in areas populated by middle- or upper-income groups. This fact is noteworthy in connection with the second topic of pertinence to the present study, namely, group-related differences in learning proficiency. It has been shown repeatedly that when learning proficiency is measured in terms of performance on standardized tests of school achievement or on commonly used tests of intelligence, children from schools serving middle- and upper-income populations are superior to children from schools serving lower-income populations (e.g., Brown & Deutsch, 1965; Wilson, 1963). It remains to be established

[1] This work was supported, in part, by a contract with the United States Office of Education (OE6-10-273) through Project Literacy. The report was prepared at the Institute of Human Learning, which is supported by grants from the National Science Foundation and the National Institutes of Health.

[2] The authors wish to thank the participating administrators, teachers, and children for their cooperation in the conduct of this study.

whether or not the deficiencies in what and how much children from low-strata schools have learned are related to concomitant deficiencies in the perfomance of such children on tasks that demand new learning. Results reported by Semler and Iscoe (1963) suggest that such a relationship may indeed hold. On a PA task, 5- and 6-year-old white children learned more efficiently than Negro children from relatively lower-strata schools. No differences were found for older children from the two groups.

One of the purposes of the present experiment was to assess the generality of the Semler and Iscoe findings for a different PA task and for groups distinguished primarily in terms of school strata rather than in terms of race. A second purpose was to determine whether or not the deficiency in PA learning expected to appear among young children from low-strata schools could be ameliorated by presenting PAs under conditions known to facilitate learning in children drawn from upper-strata schools.

METHOD

Subjects

The total sample of 432 children was drawn from three grade levels (first, third, and sixth) in two kinds of schools distinguished by the characteristic performance of their students on standardized tests of achievement and aptitude. Half the subjects (*S*s) were drawn from schools where test performance was low. Available information about the six populations from which the samples were selected is presented in Table 1. In addition to discrepancies in test performance, the high- and low-strata school populations differed in other ways associated with the distinction between "advantaged" and "disadvantaged" areas. For example, the modal occupational category of fathers of students in the high-strata schools was professional whereas that of fathers of students in the low-strata schools was semi-skilled or unskilled manual. In sum, the two populations were selected because of the contrast between them with respect to the learning proficiency of their students as assessed by standardized test performance and with respect to other characteristics often presumed to be related to the success of children in school learning.

From the total population of children within each grade level of the high- and low-strata schools, 72 *S*s were selected and assigned randomly to one or another of the six experimental

TABLE 1

MEAN CHRONOLOGICAL AGES AND STANFORD ACHIEVEMENT TEST GRADE-EQUIVALENT QUARTILES FOR GRADES 1, 3, AND 6 OF THE TWO SCHOOL-STRATA POPULATIONS

School strata	Mean CA (years)	Primary I, Form W					
		Reading					
		Q_1	Q_2	Q_3			
Grade 1							
High	6.60	1.6	1.9	2.4			
Low	6.93	1.4	1.5	1.6			
		Primary II, Form W					
		Word meaning			Paragraph meaning		
		Q_1	Q_2	Q_3	Q_1	Q_2	Q_3
Grade 3							
High	8.57						
Low	8.97	1.7	2.0	2.7	1.6	1.9	2.5
		Intermediate II, Form W					
		Word Meaning			Paragraph meaning		
		Q_1	Q_2	Q_3	Q_1	Q_2	Q_3
Grade 6							
High	11.60	5.7	6.9	8.4	6.6	7.3	8.1
Low	12.06	3.8	4.4	5.1	3.2	4.2	4.8

Note.—No data are available for the Grade 3 high-strata children.

conditions such that each cell of the design was comprised of 12 *S*s.

Materials and Design

In addition to grades and school strata, the principal factors in the 3 × 2 × 3 × 2 factorial design were Verbalization (names versus phrases versus sentences) and Depiction (still versus action). All *S*s were asked to learn the same list of 24 pairs of familiar objects presented pictorially by a pairing-test method. The three Verbalization conditions differed only with respect to the character of the experimenter's (*E*'s) utterances during the pairing trials. As each pair was presented, *E*, using a prepared script, read either the names of the two objects, (e.g., "dog. . . . gate"), a phrase containing the names of the two objects (e.g., "The dog and the gate.") or a sentence containing the names of the two objects (e.g., "The dog closes the gate."). A complete list of the verbal materials appears in Table 2. The comparison of principal interest was that between name and sentence conditions but since the presentation rate was constant for all conditions, the phrase condition was included to control for po-

TABLE 2
PAIRING-TRIAL MATERIALS

Names	Phrases/Sentences
1. Fork Cake	The fork (or/cuts) the cake.
2. Towel Plate	The towel (and/wipes) the plate.
3. Cat Log	The cat (or/jumps) the log.
4. Man Pole	The man (or/bends) the pole.
5. Bat Cup	The bat (and/strikes) the cup.
6. Shoe Chair	The shoe (and/taps) the chair.
7. Boat Ball	The boat (and/rolls) the ball.
8. Hand Hat	The hand (or/throws) the hat.
9. Rock Bottle	The rock (or/breaks) the bottle.
10. Car Wagon	The car (and/upsets) the wagon.
11. Rope Eye	The rope (and/rubs) the eye.
12. Needle Balloon	The needle (or/pops) the balloon.
13. Dog Gate	The dog (and/closes) the gate.
14. Spoon Egg	The spoon (and/rolls) the egg.
15. Fire Bed	The fire (and/burns) the bed.
16. Ax Wood	The ax (and/hits) the wood.
17. Knife Flower	The knife (or/cuts) the flower.
18. Blanket Tree	The blanket (or/covers) the tree.
19. Milk Bowl	The milk (and/fills) the bowl.
20. Teeth Apple	The teeth (or/bite) the apple.
21. Hammer Bell	The hammer (or/pulls) the bell.
22. Pencil Paper	The pencil (or/tears) the paper.
23. Doll Book	The doll (and/opens) the book.
44. Foot House	The foot (or/kicks) the house.

tential differences in rehearsal time during the pairing trials. Previous research (Rohwer, 1966) has indicated that phrases similar to those shown in Table 2 are adequate for this purpose since they do not bias performance while still providing a grammatical context to fill the pairing interval. Note that in the materials for the phrase conditions all of the forms that serve as connectives between the noun pairs are conjunctions and that only two of these are used such that each one is repeated 12 times in the list. Since there are only two repetitions of connective words in the sentence materials, it might be supposed that whatever performance differences emerge between the two conditions should be attributed to differences in intralist similarity rather than to differences in the facilitory efficacy of sentences and conjunction phrases. The results of an experiment performed to test this supposition, however, clearly indicate that it is false. Rohwer and Lynch (1967) found that sentences produce PA performance superior to that produced by conjunction phrases even when the numbers of connective repetitions were equated for the two kinds of context.

The second experimental factor, Depiction, consisted of two levels that differed with respect to whether the object pairs were presented in a manner consistent with the name and phrase verbalizations in the one case (still) or in a manner consistent with the sentence verbalizations in the other case (action). In the materials for both Depiction conditions, the pairs of objects were photographed against a background of gray cloth and their images were recorded on 16-mm. black-and-white movie film. For the still condition, the two objects in each pair were simply placed side by side on the set and photographed for 4 seconds. For the action condition, the pairs of objects were photographed while involved in the episodes described by the corresponding sentence verbalizations. The test-trial materials were the same for both Depiction conditions, that is, they consisted of 24 4-second segments of film bearing the images of the first-named objects in every pair. Two different random orders of pairing-trial and of test-trial materials were formed so that no order was repeated during the course of the two complete trials given to all *S*s.

In addition to the four principal factors in the design, the experiment was entirely balanced with respect to experimenters, of whom there were two, both male graduate students.

Procedure

The task was administered to *S*s individually for a total of two pairing and two test trials. Instructions informed the *S*s as to the procedure that would be followed and asked them to study each of the pictures of objects in order to remember which ones were presented together. One or more examples were given orally without pictorial support until *S*s appeared to understand the task. During both pairing trials, the appropriate pictorial materials were presented on a beaded screen by means of a movie projector, and, as each pair appeared, *E* read aloud the appropriate verbalization. On the test trials, *E* uttered the name of each object when it appeared on the screen and recorded *S*s' responses which were made orally. On pairing and test trials, each item was visible for 4 seconds, the interitem interval was 1 second, and the intertrial interval was 4 seconds.

RESULTS

Learning was measured in terms of the total numbers of correct responses made on the two test trials. The mean numbers of correct responses obtained by the two *E*s were very close (32.09 versus 31.61) and since a simple analysis of variance revealed that the difference was not reliable ($F < 1$) this variable was ignored in the remaining treatment of the results.

A four-way analysis of variance was applied to the data produced by the factorial design. As expected, the main effect of grades was significant ($F = 20.51$, $df = 2/396$, $p < .01$) such that the amount learned by sixth graders (33.92) was larger than that learned by third graders (32.16) which, in turn, was larger than that learned by first graders (29.48).[3] The variance associated with school strata, however, was not significant ($F < 1$); as an inspection of Table 3 suggests, the average perform-

[3] Unless otherwise indicated in the text, all post hoc comparisons were made by means of the Scheffé method at $p = .05$.

ance of children from low-strata schools was virtually the same as that of children from high-strata schools. The main effect of Verbalization was significant (F = 17.58, df = 2/396, p < .01) and a comparison of the three component means revealed that the effect was comprised entirely of the superiority of the sentence condition over both the phrase and the name conditions. Within Depiction, action was associated with more correct responses than was still (F = 108.56, df = 1/396, p < .01).

Within the analysis of variance, the tests that are critical for an evaluation of the experimental hypotheses are: The interaction of Strata × Verbalization × Depiction; and, the interaction of Grades × Strata × Verbalization × Depiction. Before the results of these tests are reported, consider the core interaction, Verbalization × Depiction, ignoring the classification variables of strata and grades. The relevant means are presented in Table 3. In agreement with our expectations, this interaction was significant (F = 14.51, df = 2/396, p < .01) and it was located entirely in the difference between those conditions designed to be facilitory and those not so designed. That is to say, the amount learned in the name-still and the phrase-still conditions was significantly smaller than the amount learned in any one of the sentence or action conditions. None of the other pair-wise contrasts was significant. Sentence verbalizations and visual translations of those sentences produced equivalent levels of performance. Similarly, name and phrase conditions were associated with equal amounts of learning, indicating that the increased opportunity for rehearsal provided in the former was of no advantage.

Turning now to the interactions of critical interest, the analysis revealed that neither was significant. The form of the interaction of Verbalization × Depiction was comparable for all *S*s, whether they were sampled from low-strata schools or from high-strata schools (F = 2.67, df = 2/396, .05 < p < .10). An examination of the means shown in Table 3 suggests, and an application of the Scheffé method confirms, that the only difference between strata lies in the marginal superiority of the low-name-action group over the high-name-action group (.05 < p < .10). Contrary to our prediction, *S*s from low-strata schools performed no less well than *S*s from high-strata schools in both the customary conditions of PA learning and in the facilitory conditions. Furthermore, the form of this three-way interaction was comparable for all grades, that is, the four-way interaction was not significant (F < 1). In sum, the present experiment produced no evidence in support of the assertion that children from low-strata schools learn PAs less efficiently than children from high-strata schools.

TABLE 3

MEAN NUMBERS OF CORRECT RESPONSES AS A FUNCTION OF STRATA, DEPICTION, AND VERBALIZATION

Depiction	School strata	Verbalization: Name	Phrase	Sentence	Total
Still	High	27.36	26.66	33.14	29.04
	Low	26.48	25.92	33.72	28.70
	Subtotal	26.92	26.29	33.43	28.88
Action	High	33.00	35.08	35.72	34.60
	Low	36.48	34.30	34.36	35.04
	Subtotal	34.74	34.69	35.04	34.82
	Total	30.83	30.49	34.24	31.85

MS_E (396) = 35.15.

The only other significant term in the analysis of variance was the interaction of Grades × Verbalization (F = 3.70, df = 4/396, p < .01). The form of this interaction was such that the difference between name and sentence conditions was larger in the first than in the third and sixth grades. The only supportable interpretation of this interaction, in view of the fact that the form of the Verbalization × Depiction interaction was equivalent for all grades (F = 1.26, df = 4/396, p > .25), is that the facilitory effects of sentence verbalization were obscured in the higher grades by the effects of the depiction conditions.

DISCUSSION

The relatively high degree of learning proficiency observed among children from low-strata schools is at once the most

puzzling and the most promising aspect of the present results. Evidence available before the learning task was administered led to the expectation that such children would be distinguished by their inability to engage successfully in new learning. If performance on standardized tests of school-related achievement is taken as an index of how much children have been able to learn up to the time of examination and if, as was the case in the present study, two groups of *S*s are equated for chronological age but still differ markedly in their test scores, a possible inference is that the members of the low-scoring group are deficient in learning ability. Obviously, this argument is too simple-minded in the sense that equivalence of chronological age does not necessarily imply equivalence of opportunity for relevant learning. Nevertheless, the teachers of the children from the low-strata schools corroborated the simplistic inference indicated by standardized test performance in describing their students as being slow to learn and difficult to teach. Furthermore, the performance of low-strata children on school-related tests of achievement is predictive of subsequent success in school learning. Thus it seems unwarranted to conclude that standardized test performance is unrelated to learning proficiency and yet the results of the present experiment seem to imply just this conclusion.

Two interpretations of the discrepancy between test and learning task performance remain to be considered. The first is that PA learning is unrelated to the kinds of learning in which a child must engage in order to perform successfully in school and on tests of school achievement. Although this interpretation cannot be discounted, we are inclined to dismiss it on the assumption that a careful description and analysis of the kind of learning teachers attempt to induce in students, especially in first-grade curricula, would reveal numerous instances of similarity to the PA paradigm. A second, and, in our view, a more likely interpretation of the discrepancy is that it occurs because of pronounced differences between the conditions of learning that are characteristic of the classroom and those that are characteristic of the laboratory.

In brief, three kinds of such differences may be distinguished. First, greater control of the focus of the child's attention is achieved in the laboratory than in the classroom by (*a*) administering the learning materials individually rather than to groups, and, in the special case of the present study, by (*b*) presenting the elements to be learned in a form that elicits the attention of the child. Second, the requirements of the child's task are explicitly detailed to a much greater extent in the laboratory than in the classroom. Third, in the laboratory case, the information necessary for the child to make a judgment about the adequacy of his performance is inherent in the learning materials themselves, whereas in the classroom such information is typically made available only in the teacher's reaction to the child's behavior and not within the boundaries of the task itself. Whether or not these differences between the conditions of learning in the classroom and in the laboratory are responsible for the discrepancy between the performance of low-strata children on standardized tests and their performance on learning tasks, it should be noted that the higher incidence of success in the laboratory than in the classroom, at least in the present study, may itself reinforce the behaviors that lead to efficient learning.

Aside from the foregoing remarks that are clearly and admittedly speculative, the results of the present experiment demonstrate empirically that the efficiency with which children, whether they are drawn from low- or high-strata schools, learn PAs can be notably affected by the manner in which the items are presented. Both the verbal condition of sentence contexts and the pictorial condition of action episodes proved facilitory for all groups of *S*s. Thus the relevance of the results to the problems of the design of educational procedures and materials is by no means confined to upper-strata populations.

The present results diverge from those

reported by Semler and Iscoe (1963) only with regard to the performance of the first-grade or 6-year old samples. This divergence may be attributable either to task differences or to differences in the way populations were defined in the two studies. More specifically, one possibility is that the method of presenting learning materials used in the present experiment may have elicited more constancy of attention from low-strata 6-year olds than that used by Semler and Iscoe. Furthermore, attention was only required for the duration of two trials in the present study in contrast to the 12 trials administered in the previous one.

REFERENCES

BROWN, B. R., & DEUTSCH, M. Some effects of social class and race on children's language and intellectual abilities: A new look at an old problem. Paper presented at the Biennial Meeting of the Society for Research in Child Development, Minneapolis, 1965.

DAVIDSON, R. E. Mediation and ability in paired-associate learning. *Journal of Educational Psychology,* 1964, **55,** 352–356.

JENSEN, A. R., & ROHWER, W. D., JR. Syntactical mediation of serial and paired-associate learning as a function of age. *Child Development,* 1965, **36,** 601–608.

REESE, H. W. Imagery in paired-associate learning in children. *Journal of Experimental Child Psychology,* 1965, **2,** 290–296.

ROHWER, W. D., JR. Constraint, syntax and meaning in paired-associate learning. *Journal of Verbal Learning and Verbal Behavior,* 1966, **5,** 541–547.

ROHWER, W. D., JR., & LYNCH, S. Semantic constraint in paired-associate learning. *Journal of Educational Psychology,* 1966, **57,** 271–278.

ROHWER, W. D., JR., & LYNCH, S. Form class and intralist similarity in paired-associate learning. *Journal of Verbal Learning and Verbal Behavior,* 1967, **6,** 551–554.

ROHWER, W. D., JR., SHUELL, T. J., & LEVIN, J. R. Context effects in the initial storage and retrieval of noun pairs. *Journal of Verbal Learning and Verbal Behavior,* 1968, in press.

SEMLER, I. J., & ISCOE, I. Comparative and developmental study of the learning abilities of Negro and white children under four conditions. *Journal of Educational Psychology,* 1963, **54,** 38–44.

WILSON, A. B. Social stratification and academic achievement. In A. H. Passow (Ed.), *Education in depressed areas.* New York: Teachers College, Columbia University, 1963. Pp. 217–235.

(Received February 7, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 1, 32–36

RELATIONSHIPS BETWEEN TRAINING METHODS AND LEARNER CHARACTERISTICS[1]

G. K. TALLMADGE[2]
American Institutes for Research in the Behavioral Sciences
Palo Alto, California

The aim of this project was to determine whether training effectiveness could be increased by employing training methods which differed as a function of trainee characteristics. A study was designed involving a control and 2 experimental training methods and 16 measures of trainee aptitudes and interests. The experimental training methods were designed to reflect Gagné's (1965) Type 3 (Chaining) and Type 7 (Principle Learning) theoretical constructs. Large achievement differences resulted from the 3 training methods. No interactions between training methods and learner characteristics were found, however, either with single aptitude measures, with combined measures, or by means of covariance analysis. It was concluded that these negative findings resulted from the existence of interactions between subject matter content and training methods.

The entire field of mental testing has grown out of an awareness that individuals differ in aptitudes, interests, and personality traits. Educational practice, however, has traditionally viewed these differences as something of an inconvenience, and has only recently recognized the potential advantages of individualized instruction.

There are several different ways in which instruction can be designed to accommodate for individual differences. To date, most research efforts in this area have been concerned with accommodating for individual differences in ability *level* and have employed such techniques as branching auto-instructional programs, self-paced learning, and others. Some research has also been conducted investigating relationships between *types* or profiles of learner aptitudes and training methods. It is this latter area which was of concern in the research reported here.

Several investigators have reported significant interactions between instructional method variables and learner characteristics but, for one reason or another, these studies have all been inconclusive. Edgerton (1958), for example, found a significant positive correlation between a word fluency test and achievement in an aircraft familiarization course taught by a "rote" method, but no correlation between the same measures when the course was taught so as to emphasize understanding. Similarly, he found a significant negative relationship between a memory test and achievement in the same course taught by the understanding method, but no relationship between these measures when the course was taught by rote methods. He concluded that trainees with high word-fluency scores should be taught by the rote method, and those with low memory scores should be taught by the "why" method. He failed to consider in this conclusion, however, either the correlation between the two aptitude measures or, more importantly, the fact that the "why" training method produced significantly higher achievement on an overall basis.

Bush, Gregg, Smith, and McBride (1965) treated the aptitude measures in

[1] This report is based on research conducted under contract N-61339-66-C-0043 with the Naval Training Device Center, Port Washington, New York and monitored by G. S. Micheli. Permission is granted for reproduction, translation, publication, use, and disposal in whole or in part by or for any purpose of the United States Government. The opinions or conclusions expressed herein are those of the author and should not be construed as necessarily reflecting the views or endorsement of the Navy Department or the Naval Training Device Center.

[2] The author gratefully acknowledges the assistance of J. W. Shearer in planning and conducting the research described herein, and of B. J. Anderson, M. A. Chapman, S. R. Ford, and J. V. Means in analyzing the data.

their study in a sound manner, but confounded training method and subject matter content in such a way that it was impossible to determine whether reported interactions were between learner characteristics and training methods or between learner characteristics and course content. The latter alternative appeared, perhaps, more likely.

Other studies could be referenced and reasons cited for questioning the apparent interactions reported between instructional methods and learner characteristics. The major difficulty appears to be that few studies have been designed specifically to assess interaction effects. Where the primary research goal has been to assess the overall effectiveness of different instructional treatments, evaluation of apparent interactions has typically been hampered by inadequate experimental or statistical controls.

The study reported here was designed specifically to test interactions between trainee characteristics (aptitudes and interests) and training methods.

METHOD

The study reported here was conducted in the setting of the United States Navy Radarman Class A (RD/A) School at Treasure Island, California. A 1-week segment of the 16-week RD/A curriculum which covered maneuvering board topics was selected as the research vehicle for the study. This choice was based on practical as well as theoretical reasons which are discussed elsewhere (Tallmadge & Shearer, 1967) and which are too lengthy to repeat here.

An analysis was made of course objectives, and statements of specific behavioral objectives were formulated. Based on these behavioral objectives, a 32-item criterion test was developed. A multiple-choice format was adopted for this test although all but four test items required some computation. The majority of test items required the plotting of points in polar coordinates, the drawing of vectors, the solution of vector diagrams, and the conversion of time, speed, and distance parameters. The incorrect answers provided for these test questions were designed to reflect common types of errors.

Two experimental versions of the 1-week maneuvering board course were designed. The first of these experimental courses (E-1) was designed to emphasize Gagné's (1965) Type 3 learning. It was oriented toward the rote memorization of fixed procedures for solving maneuvering board problems. The second experimental course was designed to reflect Gagné's Type 7 learning. It included, in addition to problem-solving procedures, the principles, concepts, and rationales underlying them. The E-2 course also involved use of a specially designed training device (Tallmadge & Shearer, 1967) to provide visual demonstrations of relevant relative motion principles.

The subjects (*Ss*) for this study were 166 Navy enlisted men enrolled at the Treasure Island RD/A School. Navy Basic Battery aptitude test scores were obtained from existing records for these *Ss*. This test battery is composed of a General Classification Test (GCT), an Arithmetic Test (ARI), a Mechanical Test (Mech), a Clerical Test (Cler), and an Electronic Technician Selection Test (ETST). The Spatial Orientation and Spatial Visualization subtests of the Guilford-Zimmerman Aptitude Survey and the Kuder Preference Record, Vocational Form B, were also administered to all *Ss*. The choice of these particular tests was based on a desire to cover all of the traditionally accepted mental abilities and to assess interests in a generic manner rather than in terms of specific vocations. A more complete description of the rationale for test selection is presented elsewhere (Tallmadge & Shearer, 1967).

Three versions of the Maneuvering Board course were included in the study: (*a*) the standard, or control, course; (*b*) E-1; and (*c*) E-2. Each version of the course was administered to an approximately equal number of trainees. The same Navy-provided instructor taught all courses throughout the study. He was intensively trained in administering the two experimental courses, and his classroom performance was continually monitored by project personnel.

Total classroom time was held constant for the three courses although substantial differences occurred in active teaching time. The E-1 course required the least teaching time and the E-2 course required the most. These differences were compensated for by providing supervised practice, review, and drill on problem-solving procedures at appropriate points throughout the course. At the end of the 1-week course, achievement was measured by means of the criterion test. Trainees were given 2 hours to complete the test.

Three alternate forms of the test were used and were administered in a random order to the 12 school classes which participated in the study. Two scores were computed for each trainee: (*a*) MB scores = items correct; and (*b*) CMB scores = items correct minus one-third items incorrect. (Note: Although all *Ss* reached the last test item, a significant number of items were omitted, particularly by the E-2 group.)

It was intended that patterns of aptitudes and interests be examined. (One plausible hypothesis was that *Ss* more interested in clerical than scientific areas would do better in the E-1 course while *Ss* with the reverse interest pattern would do better in the E-2 course.) The number of possible combi-

nations of aptitude/interest patterns was very large, however, (120 possible pairs of measures) and a decision was therefore made to examine individual measures first to look for promising leads.

For the purpose of these single measure analyses, *S*s were sorted into high and low groups on each of the 16 aptitude/interest measures. Separate two-by-three factorial design analyses of variance were conducted for each of the 16 aptitude/interest measures using each of the two dependent variables (MB scores and CMB scores). Because unequal cell frequencies resulted from the manner in which data were collected, the unweighted means technique (Winer, 1962) was used in these analyses.

RESULTS

A total of 16 unweighted means analyses of variance involving single aptitude/interest measures were computed for each of the two criterion measures. Since the results obtained with these two criterion measures were highly similar, only the MB analyses are summarized here. Table 1 presents the F ratios and associated p values resulting from the MB analyses for both main effects and the interaction effect.

TABLE 1

SUMMARY STATISTICS FROM SIXTEEN UNWEIGHTED MEANS ANALYSES OF VARIANCE USING MB CRITERION SCORES

Aptitude/interest measure	Training method	Aptitude/Interest	Interaction
	F	F	F
Navy Basic Battery			
GCT	8.95***	10.75**	<1.00
ARI	7.73***	19.09***	<1.00
Mech	8.18***	<1.00	1.04
Cler	7.38***	3.81*	1.14
ETST	6.77**	8.34**	<1.00
Guilford-Zimmerman			
GZ-V	9.85***	15.18***	<1.00
GZ-VI	11.76***	19.16***	<1.00
Kuder Vocational Preference			
K-Mech	7.38***	1.79	<1.00
K-Comp	7.79***	1.01	<1.00
K-Sci	7.12**	1.04	1.58
K-Pers	7.87***	2.81*	<1.00
K-Art	7.88***	<1.00	<1.00
K-Lit	8.09***	1.57	<1.00
K-Mus	7.62***	<1.00	<1.00
K-SocSer	7.89***	<1.00	<1.00
K-Cler	8.18***	<1.00	1.71

* $p < .10$.
** $p < .005$.
*** $p < .001$.

TABLE 2

MEANS AND STANDARD DEVIATIONS OF ACHIEVEMENT TEST SCORES FOLLOWING THREE TYPES OF TRAINING

Criterion score	Control		E-1		E-2	
	M	SD	M	SD	M	SD
MB	19.20	4.28	20.55	4.65	22.55	4.42
CMB	14.99	5.69	16.82	6.16	19.49	5.84

All analyses showed differences among training methods which were significant at either the .005 or the .001 level. Means and standard deviations of MB and CMB scores for each training method are presented in Table 2.

Five of the aptitude measures (GCT, ARI, ETST, Spatial Orientation–GZ-V, and Spatial Visualization–GZ-VI) were found to be significantly related to MB scores. These five aptitude measures were also related to CMB scores, as was the Navy Basic Battery Clerical test ($p <$ .05).

None of the analyses showed significant interactions for aptitude/interest by training method. The highest obtained F ratio was 1.71, whereas an F of 3.06 was required for significance at the .05 level (df = 2/140).

Only four of the aptitude/interest measures produced interaction F ratios greater than 1. Although these findings were not promising, mean achievement scores within cells for interaction of aptitude/interest level and training method were examined to determine whether patterns of scores on these measures could be expected to produce significant and meaningful interactions. For the purposes of this examination, control group scores were discounted since the control training method was some kind of mixture of the two experimental methods and possible interactions involving the control and one of the two experimental methods rather than the two experimental methods could not be meaningfully interpreted.

Of all the possible combinations, that involving the Navy Basic Battery Clerical test and the Kuder Scientific interest scale appeared to have the greatest potential for producing a significant interaction. The achievement difference (CMB scores) between the high and low clerical groups was 3.7 points within the E-1 training method and only 0.13 points within the E-2 training method, while scientific interest was negatively related to achievement for the E-1 method (−0.8 points) and positively related to achievement for the E-2 method (+3.0 points).

To assess this relationship statistically, the total sample of *S*s was sorted into two aptitude/interest groups: (*a*) those whose clerical achievement scores were higher than their scientific interest scores and (*b*) those whose scientific interest scores were higher than their clerical aptitude scores. CMB achievement scores were then analyzed using the same analyses of variance model previously employed for the single aptitude/interest measures. Again, the interaction effect was not statistically significant, $F = 1.12$, $df = 2/157$, although the training method main effect was significant, $F = 6.24$, $df = 2/157$, $p < .005$.

One final attempt was made to find a significant interaction between training methods and learner characteristics by employing covariance analysis techniques. The Navy Basic Battery Arithmetic test and the Guilford-Zimmerman Spatial Visualization test were selected as the most promising covariates because they were the tests most highly related to MB and CMB achievement scores. The Kuder Scientific interest scale was selected as the independent variable because it showed the highest meaningful interaction F ratio in previous analyses. (Kuder Clerical was rejected because its high interaction F ratio was caused by an uninterpretable deviation of the control group from the experimental groups in terms of the pattern of achievement scores.) MB scores were used as the dependent variable.

The results of these analyses were also negative for interaction of aptitude and method, $F = 2.44$, $df = 2/159$ with the Guilford-Zimmerman as a covariate and $F = 1.89$, $df = 2/159$ with arithmetic controlled by covariance analysis.

DISCUSSION

The main concern of the study reported here was an investigation of possible interactions existing between learner characteristics and methods of instruction. With respect to this issue, the findings of the research were negative. There were no significant interactions among the three training methods studied and the 16 learner aptitude and interest measures. In view of other reported studies, these negative results were surprising.

Although the present study cannot be considered to provide any final answers to questions about training and individual differences, it is important to seek explanations for the negative findings. The possibilities which appear most plausible are:

1. The particular training methods employed were responsible for the negative results. Other training methods might interact with learner characteristics.
2. Although the measured learner characteristics showed no interactions, other aptitude, interest, or personality factors might have.
3. Other interactions existed, perhaps between the materials to be learned and the training methods employed, which acted in such a way as to obscure the interaction of interest here.

Since this study was specifically designed to identify interactions between learner characteristics and instructional methods, and since it followed promising leads reported in other research, the first two of the above listed possibilities seemed less likely than the third. Although the study did not provide any direct support of the third alternative, its possibility was substantiated by a factor analysis (Tallmadge & Shearer, 1967) which showed the content of the Maneuvering Board course to be complex, involving manipulative skills, memory, and basic scientific background in approximately equal proportions.

Support for the existence of interactions for subject matter by training method was

provided by the study discussed earlier (Bush et al., 1965). This type of interaction could easily explain their findings. Finally, the plausibility of relationships existing between subject matter and instructional methods has led other investigators (e.g., Briggs, Campeau, Gagné, & May, 1965) to work extensively in this area.

Although this study was not designed to investigate differences in the overall effectiveness of the training methods employed, the research findings were interesting in this respect. The E-1 course was limited in content to coverage of those skills and knowledges covered by the final examination. It also provided substantially more time to drill and practice on these skills than the E-2 course. The E-2 course covered many topics not included in the criterion test, yet it produced criterion performance significantly superior to that produced by E-1 training. It was believed that this finding indicated that the E-2 course produced a higher type of learning in terms of Gagné's (1965) hierarchical structure and supported his contention tha higher types of learning are retained bette than lower types.

REFERENCES

BRIGGS, L. J., CAMPEAU, P., GAGNÉ, R. M., & MA M. A. *Instructional media: A procedure for th design of multi-media instruction, a critical re view of research, and suggestions for future re search.* Palo Alto, Calif.: American Institutes fo Research, 1965.

BUSH, W. J., GREGG, D. K., SMITH, E. A., & MCBRIDE, C. B. *Some interactions between in dividual differences and modes of instruction* Report No. AMRL-TR-65-228, Wright-Patter son AFB, Ohio: Aerospace Medical Researc Laboratories, 1965.

EDGERTON, H. A. *The relationship of method of in struction to trainee aptitude pattern.* New York: Richardson, Bellows, Henry, 1958.

GAGNÉ, R. M. *The conditions of learning.* New York: Holt, Rinehart & Winston, 1965.

TALLMADGE, G. K., & SHEARER, J. W. *Study of training equipment and individual differences, phase II.* Report No. NAVTRADEVCEN 66-C 0043-1. Port Washington, N. Y.: Naval Training Device Center.

WINER, B. J. *Statistical principles in experimental design.* New York: McGraw-Hill, 1962.

(Received February 9, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 1, 37–43

ROLE OF SPECIFIC CURIOSITY IN SCHOOL ACHIEVEMENT

HY DAY[1]
Ontario Institute for Studies in Education, University of Toronto

In a series of 3 studies the importance of specific curiosity in school achievement for Grades 7, 8, and 9 pupils was examined. A test of specific curiosity was developed based on Berlyne's definition of specific exploration. The results show that while school grades correlated significantly with IQ scores they almost invariably failed to be related with the measure of specific curiosity. The measure of specific curiosity used was shown to be related to teachers' ratings of curiosity in their pupils and to scores on the Barron-Welsh Art scale. Over an 11-month period pupils failed to increase their level of specific curiosity.

Educational systems are continually being charged with neglecting the development of curiosity in their pupils, and with rewarding students for rote learning and a display of intelligence rather than for extending their interests in the world about them. If this is so, then an examination of school grades should show that they are more closely related to IQ scores than to scores on some test of curiosity.

The concept of curiosity is of recent origin and has been poorly defined. Fowler (1965), in a recent review of this area, suggests that curiosity is "a behavior without a definition [p. 23]" although he continues to argue its existence and importance.

Maw and Maw conducted a series of studies on the evaluation of curiosity in the classroom (Maw & Maw, 1965). They defined curiosity as the need to extend one's knowledge into novel, strange, and incongruous elements in the environment. Berlyne (1963b), on the other hand, restricted his definition of curiosity or exploratory-drive to "a state of unrest and distress . . . [which] can be brought on by perceiving something under unfavourable conditions, such that the small amount of information received from the object in question leaves considerable uncertainty regarding the object's characteristics [p. 302]." This form of curiosity was said to induce the organism to engage in *specific exploration.*

It appears, therefore, that one reason for the difficulty in studying curiosity is the failure to distinguish between two types of curiosity. *Specific curiosity* describes the aroused state of an organism when confronted by an ambiguous or unclear stimulus and which may result in specific exploration; while *diversive curiosity* describes a general condition which may be analogous to what Maw and Maw consider as the need to seek new experiences or to extend one's knowledge into the unknown, and which may elicit what Berlyne has termed *diversive exploration.*

Recently it was shown that measures of exploratory choice after brief initial exposures seem to describe an inverted U-shaped function over complexity, such that there is commonly a tendency to attend to the more complex alternatives at a low level of complexity, but to avoid the more complex alternatives at a relatively higher level of complexity (Berlyne, 1963a; Day, 1965). Selective attention to more complex figures (Day, 1965) and duration of exploration (Day, 1966) also seem to increase with complexity up to a peak and then drop off. Moreover, verbal evaluation of interestingness appears to follow a similar function (Berlyne, 1963a; Day, 1965, 1967b). It seems, therefore, that the verbal response of "interesting" may indeed reflect approach behavior to complex stimulation and an intent to indulge in formation-seeking or drive-reducing exploration when in the presence of a specific complex stimulation.

It was proposed, therefore, to develop a test of specific curiosity which would

[1] The author wishes to express his gratitude to the principals and staff of the two junior high schools in the Toronto suburbs for their generous assistance in making children and facilities available.

measure the intent of the student to approach high levels of visual complexity and to withdraw from simple visual stimulation.

Using a series of 28 figures generated by Berlyne (1963a) to examine various exploratory and evaluative responses to perceptual complexity, Day had student teachers rank them along a continuum of complexity (Day, 1965). Concordance among *S*s was found to be almost perfect ($W = .91, p < .001$). Following this, other students ranked the same figures along a dimension of "interestingness." Results showed that *S*s generally tended to evaluate figures at the intermediate level of complexity as most interesting.

A Kendall test of concordance showed strong agreement among *S*s in their ranking of these figures ($W = .32, p < .001$). Yet large individual differences in shapes of the curves were found, and in the level of complexity at which their interest peaked. Vitz recently also showed that although the average shape of a preference function is an inverted U, this average is derived from a larger number of functions varying widely in shape (Vitz, 1966).

It was therefore postulated that if individual differences in preference for complex visual stimulation do, in fact, exist among school pupils, the test of specific curiosity should measure these differences. Moreover, if the charges of neglect in developing curiosity in schools today are valid, scores on this test would fail to correlate with school grades.

In a recent study, Penney and McCann (1964) found that *reactive curiosity* scores did not correlate with IQ scores in a group of 433 children in Grades 4, 5, and 6. Although *reactive curiosity* as defined by these authors appears to incorporate both specific and diversive features, it was felt that interest in complex stimulation should not of necessity be related to the ability to integrate complex stimulation into the cognitive system (certainly a feature of intelligent behavior), at least in children.

EXPERIMENT I

Method

Subjects and procedure. One hundred and thirteen pupils in Grades 7 and 8 at a junior high school in North York, Ontario, participated in the experiment. The test of specific curiosity (TSC) consisted of responses to the 28 figures projected on a screen in random order to groups of students in their classrooms, 62 subjects (*S*s) therefore being presented with the figures in one randomized order while 51 *S*s saw them in the reversed order. Each figure was displayed on a screen for a 5-second interval followed by a second 5-second interval of no-presentation to allow time for marking the answer sheets. The *S*s were instructed to evaluate their degree of interest in each figure on

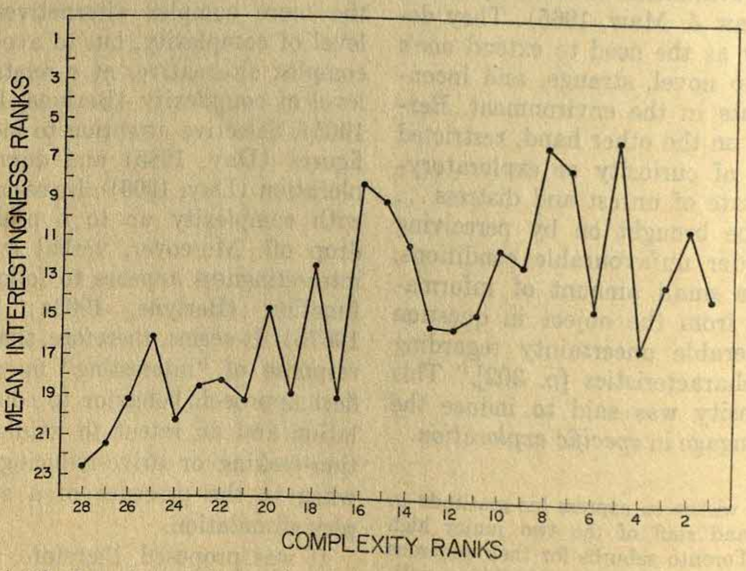


FIG. 1. Distribution of mean interestingness rankings of 28 figures ranked for complexity.

a 7-point scale. The *S*s were reassured that the information would not be made available to the school and that the purpose of the study was to evaluate the figures. Each session lasted no more than 15 minutes.

Results

Table 1 shows the results of an analysis of variance of the interestingness ratings. Order of presentation was not a significant variable but there were great differences in interest in the 28 stimulus figures. A greater portion of the total error variability appears to have been contributed by the *between Ss* variability.

IQ scores measured on the Dominion Group Test of Learning Capacity-Intermediate and end-of-term examination

TABLE 1
SUMMARY OF ANALYSIS OF VARIANCE OF INTERESTINGNESS DATA: EXPERIMENT I

Source	*df*	*MS*	*F*
Between *S*s	112	16.35	
Order	1	6.05	<1
Error (b)	111	16.45	9.73*
Within *S*s	3051	3.43	
Material	27	195.05	115.41*
O × M	27	4.65	2.75*
Error (w)	2997	1.69	

* $p < .001$.

TABLE 2
CORRELATION COEFFICIENTS OF 112 GRADE 7 AND 8 STUDENTS

Subjects	Interest in complexity	IQ
IQ	.01	—
English literature	−.01	.67*
English (language)	−.02	.64*
History	0	.61*
Spelling	.01	.59*
Math	−.02	.53*
Science	−.08	.53*
Industrial arts or home economics	−.01	.46*
Music	.02	.49*
Geography	0	.41*
Penmanship	0	.41*
Art	0	.30*

* $p < .01$.

marks in all subjects were available for 112 pupils.

A TSC score for each *S* was derived by summing the differences between the ratings of the eight most complex, and the negative differences of the six least complex figures, from the mean of that *S*'s ratings for all 28 figures. Tests of correlation of these scores with IQ scores and grade marks yielded the coefficients listed in Table 2. As predicted, TSC scores failed to correlate with IQ scores and with any of the school grades, while IQ scores corre-

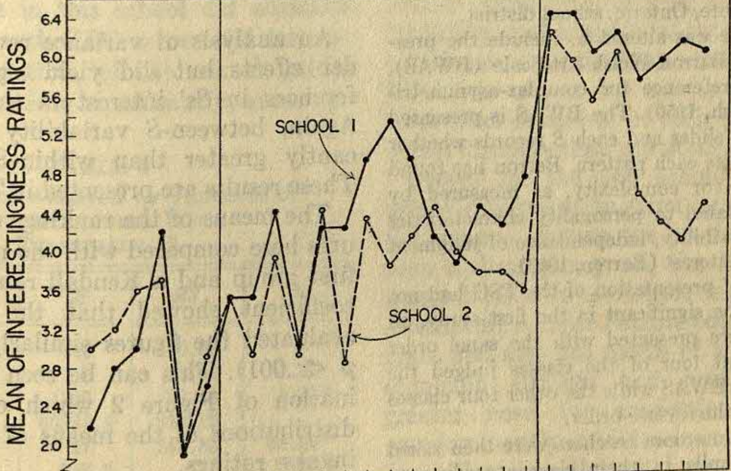


FIG. 2. Distributions of mean interestingness ratings of 28 figures ranked for complexity.

lated significantly with every one of the grade marks.

Figure 2 shows the distribution of mean interestingness ratings for the 28 figures along the dimension of complexity. The similarity of this distribution to the earlier distribution of ranking scores by adults (Figure 1) should be noted. No differences were found in specific curiosity scores between male and female students ($t = .71$).

Discussion

Interest in visual complexity appears to be unrelated to intelligence as measured by the Dominion Group Test and to grade scores, at least in this particular school. However, any generalization from these conclusions must be predicated on the findings that the test used is valid against other criteria of curiosity.

It was decided, therefore, to conduct a second study in another school in which scores on the TSC would be compared to estimations of curiosity level by the students' teachers, along the line of work done by Maw and Maw (1965).

EXPERIMENT II

Method

Subjects and procedure. The *Ss* in this study were 247 Grade 7 and 8 pupils in a junior high school in Lakeshore, Ontario, school district.

The procedure was altered to include the presentation of the Barron-Welsh-Art-Scale (BWAS), a measure of preference for complex-asymmetrical figures (Welsh, 1959). The BWAS is presented in a series of 62 slides and each *S* records whether he *likes* or *dislikes* each pattern. Barron has found that preference for complexity, as measured by this scale, is related to personality characteristics of creativity, flexibility, independence of judgment and breadth of interest (Barron, 1963).

Since order of presentation of the TSC had not been found to be significant in the first study, all eight classes were presented with the same order of the TSC, but four of the classes judged the TSC before the BWAS while the other four classes judged them in the reverse order.

The eight home-room teachers were then asked to rank the pupils in their class according to specific curiosity. In the instructions given to each teacher, specific curiosity was defined as follows:

For the purpose of this study, a school child is said to exhibit curiosity to the extent that he:

1. Shows an interest in new, incongruous, or complex objects and events in his environment.
2. Reacts positively to new, strange, incongruous, or mysterious elements in his environment by moving toward them, by exploring them, by manipulating them, or by asking questions about them.
3. Persists in examining and exploring stimuli in order to know more about them.

The following illustrations of behavior are typical of each aspect of the previous definition:

1. Shows an interest in new topics in class, asking questions or investigating them personally.
2. Reads things about strange or new subjects which are not necessarily assigned.
3. Examines, explores, and/or handles, studies, asks questions about, discusses some topic raised in class or an object brought to school by one of his classmates or the teacher.
4. Persists in such examinations, explorations, and/or manipulations (for example, he keeps studying about the topic or object until he understands it more fully).

This definition of specific curiosity was modeled on that of Maw and Maw (1965), but care was taken to extract and develop only the specific elements from their more inclusive definition.

The instructions to the teachers also suggested that the teacher list the most curious on top of the list, then the least curious on the bottom of the list, the second most curious, etc., so that the middle of the ranking would probably contain the most uncertain and ambiguous decisions.

Results

An analysis of variance revealed no order effects but did yield significant differences in *Ss'* interest in the 28 figures. Again, between-*S* variability was significantly greater than within-*S* variability. These results are presented in Table 3.

The means of the rankings of the 28 figures here compared with the means for the first group and a Kendall rank correlation coefficient showed that the two groups evaluated the figures similarly ($T = .65$; $p < .001$). This can be seen in an examination of Figure 2 which compares the distributions of the means of the interestingness ratings.

An analysis of variance of the two sets of ratings, however, showed a significant difference in interest in complexity between

the two schools. This is clearly demonstrated in Table 4. Moreover, the Schools × Material interaction was significant, indicating that the differences were in the slopes of the distributions.

An examination of Figure 2 shows that the differences occur mainly in the extremes of the distribution, outside of the two vertical lines, and including the data which are used to establish the specific curiosity scores. A t test of differences in curiosity scores yielded highly significant results ($t = 6.49$; $p < .001$), showing that the pupils in the first school had a generally higher level of specific curiosity.

Again, sex differences were not significant ($t = .081$).

Since most of the classes numbered 30–35 pupils it was decided to compare scores of the pupils ranked by the teacher in the top quarter of the class (eight *Ss*) with those in the bottom quarter. Results of a t test of differences was highly significant ($t = 12.6$; $p < .001$), indicating that the TSC was in substantial agreement with the teachers' evaluations of curiosity in their pupils.

TSC scores correlated significantly with BWAS scores ($r = .22$; $p < .01$), suggesting that high curiosity *Ss* tended to be flexible and independent of judgment. Intercorrelations with school marks are presented in Table 5. They showed that some of the grades in this school did correlate significantly with TSC scores. Of interest is the finding that, of the school grades which did correlate significantly, some appear to lack any relationship with curiosity (e.g., penmanship).

TABLE 3

SUMMARY OF ANALYSIS OF VARIANCE OF INTERESTINGNESS DATA: EXPERIMENT II

Source	*df*	*MS*	*F*
Between *Ss*	246	10.96	
Order	1	36.62	3.37
Error (b)	245	10.86	4.23*
Within *Ss*	65335.5	248.20	
Material	27	245.53	95.55*
O × M	123.5	.10	.04
Error (w)	65185	2.57	

* $p < .001$.

TABLE 4

ANALYSIS OF VARIANCE OF INTERESTINGNESS DATA BETWEEN THE TWO SCHOOLS

Source	*df*	*MS*	*F*
Between *Ss*	359	561.91	
Schools	1	544.36	31.0*
Error(b)	358	17.55	7.6*
Within *Ss*	9720	4.01	
Material	27	391.33	170.1*
S × M	27	228.83	99.5*
Error(w)	9666	2.30	

* $p < .001$.

IQ scores were not available at this school, so a correlation of specific curiosity with intelligence was not obtained.

DISCUSSION

The results of these two studies point clearly to the existence of a dimension of specific curiosity which can be defined as an "interest in complex stimulation." The test developed to measure this dimension appears to discriminate levels of specific curiosity, at least among adolescents between the ages of 12 and 16. Moreover, scores on this test were substantially in agreement with teachers' ratings of their pupils along this dimension, and correlated significantly with Barron's description of individuals varying along a trait of flexibility and breadth of interest. The low correlation may be due to the fact that the BWAS is a test for adults who may be more discriminating in their preference for visual displays than are adolescents.

Scores for the two schools are significantly discrepant to stimulate a search for the reasons. Among the possible causes for such a discrepancy are the facts that the first school is in a district which has been attempting a more flexible educational program and also has pupils of a higher socioeconomic stratum and possibly with a greater need for academic achievement (predominantly middle-class Jewish families). However, the scope of this study did not include an examination of the interschool differences in curiosity.

TABLE 5

Correlation Coefficients of Specific Curiosity and School Grades of 247 Grade 7 and 8 Students

Subjects	r
English literature	.029
English composition	.224**
English grammar	−.001
History	.026
Mathematics	.030
Physical education	.009
Geography	.060
Science	.172*
Penmanship	.170*
Music	.158*
Spelling	.428**
Industrial arts or home economics	.239**
Barron-Welsh Art Scale	.220**

* $p < .05$.
** $p < .01$.

Experiment III

Method

Subjects and procedure. The *Ss* in this study were 429 Grade 7, 8, and 9 pupils in the same junior high school as in Experiment I.

The TSC was presented in the school auditorium with a complete grade participating in a session. The BWAS was also presented at the same sessions.

Results

Of the 113 pupils who had taken the test in the previous sample, only 61 were present for the retest. Test-retest correlation was .48, significant at the 1% level. A t test failed to show any significant shift in specific curiosity scores over the 11-month period ($t = .66$).

The distributions of mean ratings of interest-in-complexity scores and TSC scores were not significantly different from those in Experiment I.

Correlation coefficients again showed that IQ scores correlated significantly with school grades (except penmanship) but not with TSC scores ($r = .01$). On the other hand TSC scores correlated significantly only with English composition, penmanship, and music ($p < .05$). TSC scores again correlated with BWAS scores ($r = .14$, $N = 428$, $p < .01$). Table 6 summarizes these correlation coefficients.

Unlike Penney and McCann's (1964) findings, there were no significant sex differences in curiosity. However, this is in line with Maw and Maw (1964), who reported no consistent sex differences in the selection of unbalanced and/or unfamiliar symbols and figures.

Discussion

The importance of specific curiosity as a dimension in a student's behavior need hardly be argued. A high level of specific curiosity indicates an interest in approaching and exploring high levels of novelty, complexity, incongruity, etc. Certainly such a characteristic mode of behavior should lead to the development of an individual who seeks to learn and to develop. Barron (1963) and Golan (1962) showed that preference for asymmetry and complexity in visual patterns is linked with high creativity and flexibility. If one goal of education is the development of creative individuals, then the educational system should inspire and reward curiosity.

However, results in this series of experiments indicate the failure of the present educational system to develop interest in complexity. This is demonstrated in the failure of students to improve scores on a test of specific curiosity over 11 months and in the failure to find a consistent cor-

TABLE 6

Correlation Coefficients of Specific Curiosity and IQ Scores of Grade 7, 8, and 9 Students

Subjects	Specific curiosity	N	IQ	N
IQ	−.01	395	—	—
English literature	.01	427	.40**	439
English (composition)	.09*	427	.39**	440
Spelling	.09	300	.24**	309
Penmanship	.13*	300	.03	309
History	−.03	428	.28**	440
Geography	.00	429	.37**	441
Mathematics	−.06	426	.45**	439
Science	.00	429	.44**	441
French	.04	416	.40**	428
Music	.10*	348	.32**	360
Art	.02	314	.19**	323
Industrial arts or home economics	.03	342	.27**	352
Typing	.08	302	.17**	310
Physical and health education	−.04	429	.24**	441
Barron-Welsh Art Scale	.14**	428	.02	394

* $p < .05$.
** $p < .01$.

relation between school grades and TSC scores. Examination of the school grade–curiosity correlations shows that no subject scores are consistently correlated with TSC scores although a few of them (composition, penmanship, and music) did correlate significantly in two of three studies.[2] It is interesting to speculate on the reasons for these correlations since these three subjects emphasize orderliness, symmetry, and aesthetic qualities, but any answers to this question would require additional studies directly concerned with an understanding of curriculum content in the various subjects.

On the other hand, intelligence seems to be an important correlate of school achievement accounting for a fair proportion of variance. This concurs with the commonly accepted notion that achievement in the classroom requires intelligent behavior but little or no interest in expanding one's knowledge into areas peripherally or indirectly connected with curriculum.

Further work is presently in progress which will relate the concept of specific curiosity to that of anxiety and to other personality characteristics in adolescents (Day, 1966). A series of studies has also been undertaken with the goal of examining the methods of manipulating the level of specific curiosity (Day, 1967a; Day & Thomas, 1967; Sobol & Day, 1967). Results to date emphasize the importance of maintaining arousal level at an optimum level in order to achieve maximum curiosity.

REFERENCES

BARRON, F. Discovering the creative personality. In *College admission I.Q.: The behavioral sciences and education.* New York: College Entrance Examination Board of New York, 1963. Pp. 3–7.

BERLYNE, D. E. Complexity and incongruity variables as determinants of exploratory choice and evaluative ratings. *Canadian Journal of Psychology,* 1963, **17**, 274, 290.(a)

BERLYNE, D. E. Motivational problems raised by exploratory and epistemic behaviour. In S. Koch (Ed.), *Psychology: A study of a science,* New York: McGraw-Hill, 1963. Pp. 284–364.(b)

DAY, H. Exploratory behaviour as a function of individual differences and level of arousal. Unpublished doctoral dissertation, University of Toronto, 1965. (Microfilm #66-1064, University Microfilms, Ann Arbor, Michigan.)

DAY, H. Looking time as a function of stimulus variables and individual differences. *Perceptual and Motor Skills,* 1966, **22**, 423–428.

DAY, H. The effects of increased arousal on attention in high- and low-anxious subjects. *Ontario Journal of Educational Research,* 1967, **9**, 185–191.(a)

DAY, H. Evaluations of subjective complexity, pleasingness and interestingness for a series of random polygons varying in complexity. *Perception and Psychophysics,* 1967, **2**, 281–286.(b)

DAY, H., & THOMAS, L. E. Effects of amphetamine on selective attention. *Perceptual and Motor Skills,* 1967, **24**, 1119–1125.

FOWLER, H. *Curiosity and exploratory behavior.* New York: Macmillan, 1965.

GOLAN, S. E. The creativity motive. *Journal of Personality,* 1962, **30**, 588–600.

MAW, W. H., & MAW, E. W. The measurement of curiosity in elementary school children. A summary. *Cooperative Research Project No. 801.* Moravia, N.Y.: Chronicle Guidance Publications, 1964.

MAW, W. H., & MAW, E. W. Personal and social variables differentiating children with high and low curiosity. Cooperative Research Project No. 1511, University of Delaware, 1965.

PENNEY, R. K., & MCCANN, B. The children's reactive curiosity scale. *Psychological Reports,* 1964, **15**, 323–334.

SOBOL, M., & DAY, H. The effect of colour on exploratory behavior and arousal. Paper read at the 28th Annual Meeting of the Canadian Psychological Association, Ottawa, June 1967.

VITZ, P. E. Preference for different amounts of visual complexity. *Behavioral Science,* 1966, **2**, 105–114.

WELSH, G. S. *Welsh Figure Preference Test, preliminary manual.* Palo Alto: Consulting Psychologists Press, 1959.

(Received February 13, 1967)

[2] In a recently completed study of 188 pupils TSC scores failed to correlate significantly with any school grades.

Journal of Educational Psychology
1968, Vol. 59, No. 1, 44–52

STUDENT PERSONALITY CORRELATES OF TEACHER RATINGS

GEORGE D. YONGE AND JULIUS M. SASSENRATH
University of California, Davis

The 227 students taught by 3 instructors of a course in educational psychology completed an instructor/course rating form and a personality test. The ratings were factored and 9 factors were rotated by the normalized varimax technique. Factor scores were computed for each student. Correlations between the personality test and factor scores were computed separately for each of the 3 instructor samples. Differences between correlations were tested for significance among the 3 instructors. The results indicate that generally different personality scores correlate with a given factor score, and in some instances the same personality variable correlates with a given factor score in opposite directions from one instructor to another. It was concluded that the factors have a different psychological meaning for each instructor. The results were discussed in terms of the known differences among the instructors.

Of the many studies dealing with teacher ratings, only a few have considered the personality correlates of these evaluations (Getzels & Jackson, 1963; McKeachie, 1963; Remmers, 1963). Further, of these few studies most have been concerned with correlations between ratings and the personality characteristics of the teachers. The results are consistent enough to indicate that ratings are correlated with measured aspects of an instructor's personality obtained by self report and peer nomination techniques (Corcoran, 1961; Isaacson, McKeachie, & Milholland, 1963; Veldman & Peck, 1963). Evaluations of instructors, then, appear to reflect something more than a student's personal reactions.

An equally important but somewhat neglected aspect of teacher rating is their relationship to the personality characteristics of those doing the rating (Rezler, 1965). That individuals with different education values describe different traits as being generally important for teachers is not surprising (Kerlinger, 1966); any evaluative decision or choice must be viewed in terms of a preexisting frame of reference from which the value or weight of a given dimension is derived. It is contended that certain measured personality characteristics of students reflect to some extent these preexisting frames of reference. These frames of reference are not merely something brought to the classroom by the student, but rather they are part of the total context in which the instructor's behavior occurs. As part of the context, even if only an implicit part of the background, they contribute to the meaning of an instructor's behavior or teaching.

Where Rezler (1965) studied student personality correlates of teacher ratings, ignoring instructor differences, the purpose of the present study is to show that student personality correlates of teacher ratings vary markedly from one instructor to another when students of similar average personality characteristics rate these instructors with the "same" evaluation form.

METHOD

Instructor Sample

Three instructors of a course in introductory educational psychology were rated on several scales by students enrolled in this course during the fall and spring semesters of 1965–1966. The course is essentially a lecture course. All sections use a common textbook, and one half of each examination (covering the textbook) is common. Each instructor is free to develop the lecture aspect of the course in whatever way he deems appropriate. Instructor A has chosen to emphasize the classroom application of psychological facts, and he also makes greater use of audio-visual aids than Instructors B and C. Instructor B emphasizes the psychological aspect of the subject matter with some explicit attempt to develop the educational implications of experimental results in psychology and education. Instructor C empha-

sizes a humanistic psychology focusing on topics of relevance to education but with little explicit development of classroom implications.

The three instructors may be ranked in terms of an applied-theoretical course emphasis with A being most applied and C most theoretical. Since both B and C offer primarily a lecture course, their style of teaching is more similar to each other than to A, although in terms of the purpose of the course offered, A and B are more similar. The current interests of the three instructors are also somewhat diverse. Instructor A, by training the interest, is oriented toward guidance and counseling, Instructor B is interested in experimental and methodological procedures, and Instructor C is interested in certain philosophical aspects of psychology. To some extent these interests are probably expressed in the way each teaches the course.

Student Sample

Both the evaluation form and a personality test were completed by 227 students. This represents approximately 70% of the total course enrollment. There are 50 men and 177 women in the sample, most being in their junior or senior year. Since for each instructor the patterns of correlations among ratings and between ratings and personality test scores are similar for men and women, as well as for both semesters, these data have been combined for each of the three teachers. Instructor A, who taught two sections of the course, was rated by a total of 39 students, Instructor B, who taught four sections of the course, was rated by a total of 120 students, and Instructor C, who taught three sections, was rated by a total of 68 students who also completed the personality test. The average scores of students taught by the three instructors do not differ significantly on any of the 14 personality measures nor are there significant differences among instructors on each of the nine factor scores.

Measures

The evaluation form contains 21 items dealing with teacher behavior and 12 focusing on aspects of the course. These items were presented as seven-point rating scales and were adapted from published and unpublished forms, or were written by the present investigators.

Thirty-two[1] of the items from the evaluation form were factored by means of an incomplete components analysis using ones in the diagonal of the intercorrelation matrix. This analysis was performed on the zero order correlations derived from the total sample of 227 students. Nine factors were selected for rotation by the normalized varimax technique, and factor scores were computed for each student on the nine factors. The items on the evaluation form which loaded .40 or higher on the nine factors are presented in Table 1.

TABLE 1

FACTOR STRUCTURE OF THE EVALUATION FORM FOR 227 STUDENTS (LOADINGS OF .40 OR GREATER)

Factor	Loading
Factor I: Confident, fluent delivery	
Self-reliance and confidence	.74
Speaks clearly and distinctly	.71
Speaks fluently and without hesitation	.71
Free from annoying mannerisms	.57
Permissive and flexible	.41
Factor II: Clarity of course	
Objectives of course clear	.83
Course well organized in meaningful sequence	.82
Assignments clear and reasonable	.48
Explanations clearly presented	.44
Factor III: Open, sympathetic attitude toward students	
Willingness to help students	.73
Students feel free to express own opinions	.70
Sympathetic attitude toward students	.70
Respectful of views other than own	.64
Open to comments and questions	.61
Sense of proportion and humor	.60
Permissive and flexible	.49
Free from annoying mannerisms	.47
Factor IV: Interesting, stimulating teacher	
Stimulates intellectual curiosity	.79
Presents material in interesting way	.76
Relates subject to other fields	.70
All-around teaching ability	.67
Factor V: Fairness of evaluation	
Content of exams appropriate	.82
Fairness of grades	.80
Factor VI: Suitable class material and value of course	
Contribution of course to general education	.79
Value of course for teacher preparation	.78
No overlap with other courses taken	.74
Materials suited to class level	.70
Factor VII: Frequency of evaluation	
Frequency of papers adequate	.74
Frequency of exams appropriate	.74
Factor VIII: Interest in and knowledge of subject matter	
Knowledge of subject matter	.78
Interest in subject	.74
Knowledge of new developments in field	.68
Factor IX: Preparedness of lectures	
Classes well prepared	.66
Uses appropriate techniques of instruction	.45
Explanations clearly presented	.42

In addition to the above listed items and the resulting factor scores, semester grade (Gr) in the course was another variable. Further, all students were given the Omnibus Personality Inventory (OPI), Form F (Center for the Study of Higher Education, 1962). The OPI was derived explicitly for use with college students. It yields 14 measures of relevance to an academic context. Brief scale descriptions follow.

Thinking Introversion (TI). Liking for reflective thought and scholarly activities.

Theoretical Orientation (TO). Interest in science and scientific activities.

Estheticism (Es). Diverse interests in artistic matters and activities.

Complexity (Co). Tolerance for ambiguities and uncertainties, a fondness for novel situations and ideas.

Autonomy (Au). Nonauthoritarianism and a need for independence.

[1] The item dealing with the adequacy of the textbook was omitted because different texts were used from one semester to the other.

Religious Orientation (RO). Skeptical of orthodox religious beliefs and practices.

Social Extroversion (SE). Strong interest in people and being with them.

Impulse Expression (IE). Readiness to express impulses and to seek gratification in thought and action.

Personal Integration (PI). Admits to few attitudes and behaviors that characterize socially alienated persons.

Anxiety Level (AL). High scorers deny feelings or symptoms of anxiety, and do not admit to being nervous or worried.

Altruism (Am). Strong concerns for the welfare and feelings of people.

Practical Outlook (PO). Evaluates ideas and things in terms of immediate utility; values material possessions and concrete accomplishments.

Masculinity-Femininity (MF). High scorers (masculine) deny interests in esthetic matters, admit to few adjustment problems, express an interest in scientific matters.

Response Bias (RB). High scorers are responding to this measure in a manner similar to a group of students who were explicitly asked to make a good impression on the test.

A 47 by 47 variable (14 OPI, 32 ratings, grade) correlation matrix was calculated separately for the students of each instructor. Some of these correlations will be presented below. In addition, the nine factor scores were correlated with the OPI scores and grades for the students of each instructor and form the major basis of the present study.

RESULTS

Teacher Factors

The data presented below show the personality and grade correlations with the scores for the five factors pertaining to teacher behavior. Not only are there noticeable differences in the direction of the correlations from one instructor to another, but there are several significant differences between correlations.

Although the primary purpose of the present study is focused on differences in correlations from instructor to instructor, it is also instructive to note the types of student who are differentially appealed to by these instructors. Many of the results to be presented can be understood in the light of what was said previously about each instructor.

The significant personality correlates of Factor I, confident, fluent delivery, for the three instructors are presented in Table 2.

TABLE 2

CORRELATES OF FACTOR I, CONFIDENT, FLUENT DELIVERY, SIGNIFICANTLY DIFFERENT FROM ZERO WITHIN AND/OR SIGNIFICANTLY DIFFERENT AMONG INSTRUCTOR SAMPLES

Correlates	Instructor A	Instructor B	Instructor C	Significantly different correlations[a]
Es	−.19	−.20*	.29*	AB versus C
PI	.35*	.08	.15	
AL	.32*	.13	−.05	
PO	.09	.20*	−.09	
MF	.33*	.18*	−.12	A versus C
RB	.43**	.09	.09	
Gr	.06	.01	.33**	B versus C

Note.—For Instructor A, $N = 39$; for Instructor B, $N = 120$; for Instructor C, $N = 68$.

[a] Differences between Z transformed correlations exceed twice the standard error of the difference. For example, with respect to Es, the correlation based on the sample for Instructor A and for Instructor B is each significantly different from the correlation based on the sample for Instructor C.

* $p < .05$.

** $p < .01$.

As may be noted, there are significant differences in the personality test correlations among instructor samples on the measures of esthetic interests, masculine-feminine interests, and semester grade. As indicated by the significant correlates of this factor for each instructor, a different type of student rates each instructor high on confident, fluent delivery of lectures. Specifically, the significant correlates for Instructor A are Response Bias (.43), Personal Integration (.35), Masculinity (.33), and Anxiety Level (.32). Thus, students rating Instructor A high on this factor may be characterized as having a positive self-regard and masculine interests. The significant correlates for Instructor B are Practical Outlook (.20), Estheticism (−.20), and Masculinity-Femininity (.18), thus characterizing students who rate Instructor B high as being practical, nonesthetic, and masculine in interests. On the basis of the significant correlates of this factor for Instructor C, students giving this instructor a high rating may be characterized as esthetically oriented (Es .29) with a high level of achievement in the course (Gr .33).

The personality measures related to Factor III, open, sympathetic attitude toward students, for the three instructors are presented in Table 3.

There are significant differences in correlations among the three instructors on the measures of Estheticism, Social Extrover-

sion, and Anxiety Level. The significant correlates of this factor for Instructor A, Anxiety Level (.35), Personal Integration (.34), Response Bias (.34), and Altruism (.32), indicate that the type of student who perceives this instructor as open and sympathetic is the student with a positive self-regard and a concern for the welfare of others. The practically oriented (PO .19) student tends to rate Instructor B high, and the esthetically oriented (Es .29) student tends to rate Instructor C high on this factor.

The significant personality correlates of Factor IV, interesting, stimulating teacher, for the three instructors are to be found in Table 4.

Nine of these 10 variables show significant differences in correlations among the three instructors. With respect to the type of student rating each instructor high on this factor, the correlates for Instructor A, Personal Integration (.44), Response Bias (.43), Complexity (−.37), and Anxiety Level (.36), indicate that students with a positive self-regard and who like well-structured, unambiguous situations rate him high. The significant correlates for Instructor B, Anxiety Level (.23), Religious Orientation (−.23), Masculinity-Femininity (.18), and Response Bias (.18) suggest the characterization of a student with a positive self-regard, with strong religious be-

TABLE 3

CORRELATES OF FACTOR III, OPEN, SYMPATHETIC, SIGNIFICANTLY DIFFERENT FROM ZERO WITHIN AND/OR SIGNIFICANTLY DIFFERENT AMONG INSTRUCTOR SAMPLES

Correlates	Instructor A	Instructor B	Instructor C	Significantly different correlations[a]
Es	−.07	−.12	.29*	B versus C
SE	.31	−.12	−.06	A versus B
PI	.34*	.10	.06	
Al	.35*	.12	−.08	A versus C
Am	.32*	.10	.20	
PO	−.06	.19*	−.10	
RB	.34*	.11	.00	

Note.—For Instructor A, $N = 39$; for Instructor B, $N = 120$; for Instructor C, $N = 68$.

[a] Differences between Z transformed correlations exceed twice the standard error of the difference. For example, with respect to Es, the correlation based on the sample for Instructor B is significantly different from the correlation based on the sample for Instructor C.

* $p < .05$.

** $p < .01$.

TABLE 4

CORRELATES OF FACTOR IV, INTERESTING, STIMULATING TEACHER, SIGNIFICANTLY DIFFERENT FROM ZERO WITHIN AND/OR SIGNIFICANTLY DIFFERENT AMONG INSTRUCTOR SAMPLES

Correlates	Instructor A	Instructor B	Instructor C	Significantly different correlations[a]
Es	−.23	−.15	.37**	AB versus C
Co	−.37*	−.09	.23*	A versus C
Au	−.16	−.11	.21	B versus C
RO	.09	−.23**	−.06	
PI	.44**	.14	.06	A versus C
AL	.36*	.23**	−.07	AB versus C
PO	.26	.16	−.17	AB versus C
MF	.24	.18*	−.14	B versus C
RB	.43**	.18*	.03	A versus C
Gr	−.00	.12	.41**	AB versus C

Note.—For Instructor A, $N = 39$; for Instructor B, $N = 120$; for Instructor C, $N = 68$.

[a] Differences between Z transformed correlations exceed twice the standard error of the difference. For example, with respect to Es, the correlation based on the sample for Instructor A and for Instructor B is each significantly different from the correlation based on the sample for Instructor C.

* $p < .05$.

** $p < .01$.

liefs, and masculine interests as rating him high on Factor IV. The three correlates for Instructor C, Grades (.41), Estheticism (.37), and Complexity (.23), indicate that students perceiving Instructor C as an interesting stimulating teacher may be characterized as being high achievers in the course, as having esthetic interests, and as liking ambiguous, unstructured situations as well as novel ideas. The fact that Complexity correlates significantly with this factor for both Instructors A and C but in *opposite* directions is worthy of note because it points clearly to the very thesis of the present paper; namely, the psychological meaning of each factor varies across instructors.

The personality measures related to Factor VIII, interest in and knowledge of subject matter, for the three instructors are presented in Table 5.

As indicated, seven of the personality measures have significantly different correlations with this factor across the three instructors. The measures of Personal Integration (.37), Response Bias (.37), Complexity (−.37), and Anxiety Level (.35), indicate that the type of student rating Instructor A high on this factor may be characterized as having a positive self-

TABLE 5

CORRELATES OF FACTOR VIII, INTEREST IN AND KNOWLEDGE OF SUBJECT MATTER, SIGNIFICANTLY DIFFERENT FROM ZERO WITHIN AND/OR SIGNIFICANTLY DIFFERENT AMONG INSTRUCTOR SAMPLES

Correlates	Instructor A	Instructor B	Instructor C	Significantly different correlations[a]
TI	−.19	−.19*	.15	B versus C
Es	−.21	−.15	.19	B versus C
Co	−.37*	−.17	.14	AB versus C
Au	−.24	−.16	.15	B versus C
RO	−.00	−.22*	.03	
SE	.29	−.11	.02	A versus B
PI	.37*	.10	.09	
AL	.35*	.18*	−.04	A versus C
PO	.28	.22*	−.16	AB versus C
RB	.37*	.09	.10	

Note.—For Instructor A, $N = 39$; for Instructor B, $N = 120$; for Instructor C, $N = 68$.

[a] Differences between Z transformed correlations exceed twice the standard error of the difference. For example, with respect to Co, the correlation based on the sample for Instructor A and for Instructor B is each significantly different from the correlation based on the sample for Instructor C.

* $p < .05$.

** $p < .01$.

regard, and as liking structured, unambiguous situations. The correlates for Instructor B, Practical Outlook (.22), Religious Orientation (−.22), Thinking Introversion (−.19), and Anxiety Level (.18), indicate that students rating this instructor high tend to be practical, religious, nontheoretical, and nonanxious. None of the personality correlates for Instructor C are significantly different from zero.

The significant personality correlates of the final teacher dimension, Factor IX, preparedness of lectures, for the three instructors are to be found in Table 6.

Five personality measures show significantly different correlations with this factor among the three instructors. This is the first factor encountered in which the measures of personal, social adjustment (positive self-regard) do not correlate for Instructor A. Rather, the type of student rating him high has masculine interests (MF .33), likes well-structured, unambiguous situations (Co −.33), and does not have esthetic interests (Es −.32). The one personality measure correlated with this factor for Instructor B indicates that the practically oriented student (PO .18) is likely to perceive him as having well-prepared lectures. The personality correlates for Instructor C allow a familiar characterization of the type of student who rates him as having well-prepared lectures—a high achiever in the course (Gr .36), with esthetic (Es .28) and feminine (MF −.23) interests.

Course Factors

The personality measures related to the four factors describing aspects of the course will now be presented. The student personality correlates of Factor II, clarity of course, for the three instructors are shown in Table 7.

There are more significantly different personality correlates (10) of this factor across the three instructors than for any other course or teacher dimension. Further, for Instructor A, this is the factor with the greatest number of significant correlates. The six personality correlates for Instructor A are Personal Integration (.50), Anxiety Level (.48), Response Bias (.45), Impulse Expression (−.39), Social Extroversion (.37), and Complexity (−.32). Hence, the characterization of the student rating his course high on this factor is one having a positive self-regard, as being nonimpulsive, socially extroverted, and preferring well-structured, unambiguous situations. The three correlates of this factor for Instructor B indicate that students rating his course high in clarity are religiously oriented (RO −.24), nonanxious (AL .23),

TABLE 6

CORRELATES OF FACTOR IX, PREPAREDNESS OF LECTURES, SIGNIFICANTLY DIFFERENT FROM ZERO WITHIN AND/OR SIGNIFICANTLY DIFFERENT AMONG INSTRUCTOR SAMPLES

Correlates	Instructor A	Instructor B	Instructor C	Significantly different correlations[a]
Es	−.32*	−.13	.28*	AB versus C
Co	−.33*	−.15	.12	A versus C
PO	.25	.18*	−.16	AB versus C
MF	.33*	.10	−.23*	AB versus C
Gr	−.09	.00	.36**	AB versus C

Note.—For Instructor A, $N = 39$; for Instructor B, $N = 120$; for Instructor C, $N = 68$.

[a] Differences between Z transformed correlations exceed twice the standard error of the difference. For example, with respect to Es, the correlation based on the sample for Instructor A and for Instructor B is each significantly different from the correlation based on the sample for Instructor C.

* $p < .05$.

** $p < .01$.

and do not admit to personal and psychological shortcomings (RB .21). With respect to Instructor C, high achievers in the course rate it high in clarity (Gr .36).

Factor V, fairness of evaluation, reflects one of the structural aspects of the course in that all three instructors used essentially the same procedures (examinations and a paper) and criteria for evaluating the students. As might be expected, the personality correlates of this factor are minimal. In fact, only two variables correlate significantly at the .05 level with this factor. These are Anxiety Level for Instructor A and Grades for Instructor C. The Anxiety Level correlation based on the sample for Instructor A (.36) is significantly different from the correlation based on the sample for Instructor C (−.08). The correlate for Instructor A suggests that the less anxious a student the higher he will rate the fairness of evaluation. For Instructor C, the better the course grade, the more favorable the rating on this factor. There are no significant correlates for Instructor B.

The significant personality correlates of Factor VI, suitable class material and value of course, for the three instructors are presented in Table 8.

Eight variables showed significant differences in correlations among the three instructors. The measures having a significant relationship for Instructor A indicate that students rating his course favorably do not like ambiguous, uncertain situations (Co −.40), they do not admit to feelings of anxiety (AL .39), they are socially outgoing (SE .38), and they are nonimpulsive (IE −.32). The type of student perceiving the course given by Instructor B as high is practical in outlook (PO .24), religious (RO −.24), authoritarian (Au −.19), uninterested in reflective, theoretical thinking (TI −.18), and not esthetically inclined (Es −.18). Finally, for Instructor C, high achievers in the course tend to perceive the materials as suitable and the course of value (Gr .41).

With respect to the final course dimension, Factor VII, frequency of evaluation, it is not surprising that none of the personality variables correlate significantly with this factor. This dimension reflects a structural feature of the course which is constant across all three instructors—the same number of examinations and papers are required. It is, however, interesting to note that the correlations between this factor and Autonomy (−.13, .18) as well as Practical Outlook (.16, −.17) are significantly different for Instructors B and C.

TABLE 7

CORRELATES OF FACTOR II, CLARITY OF COURSE, SIGNIFICANTLY DIFFERENT FROM ZERO WITHIN AND/OR SIGNIFICANTLY DIFFERENT AMONG INSTRUCTOR SAMPLES

Correlates	Instructor A	Instructor B	Instructor C	Significantly different correlations[a]
Co	−.32*	−.10	.12	A versus C
Au	−.16	−.10	.18	B versus C
RO	−.08	−.24**	−.09	
SE	.37*	−.01	−.05	A versus BC
IE	−.39*	−.07	.03	A versus C
PI	.50**	.15	.04	A versus BC
AL	.48**	.23**	−.13	AB versus C
PO	.14	.12	−.21	B versus C
MF	.22	.15	−.19	AB versus C
RB	.45**	.21*	.04	A versus C
Gr	.14	.06	.36**	B versus C

Note.—For Instructor A, $N = 39$; for Instructor B, $N = 120$; for Instructor C, $N = 68$.

[a] Differences between Z transformed correlations exceed twice the standard error of the difference. For example, with respect to AL, the correlation based on the sample for Instructor A and for Instructor B is each significantly different from the correlation based on the sample for Instructor C.

* $p < .05$.

** $p < .01$.

TABLE 8

CORRELATES OF FACTOR VI, SUITABLE MATERIAL AND COURSE VALUE, SIGNIFICANTLY DIFFERENT FROM ZERO WITHIN AND/OR SIGNIFICANTLY DIFFERENT AMONG INSTRUCTOR SAMPLES

Correlates	Instructor A	Instructor B	Instructor C	Significantly different correlations[a]
TI	−.26	−.18*	.05	
Es	−.24	−.18*	.18	AB versus C
Co	−.40**	−.16	.11	A versus C
Au	−.23	−.19*	.18	AB versus C
RO	−.10	−.24*	.08	B versus C
SE	.38*	−.04	−.08	A versus BC
IE	−.32*	−.01	−.01	
AL	.39*	.14	−.03	A versus C
PO	.27	.24*	−.18	AB versus C
Gr	.05	−.04	.41**	B versus C

Note.—For Instructor A, $N = 39$; for Instructor B, $N = 120$; for Instructor C, $N = 68$.

[a] Differences between Z transformed correlations exceed twice the standard error of the difference. For example, with respect to Es, the correlation based on the sample for Instructor A and for Instructor B is each significantly different from the correlation based on the sample for Instructor C.

* $p < .05$.

** $p < .01$.

These differences in correlations probably reflect the fact that these structural constants of the course are viewed in a slightly different way across instructors because these aspects are experienced as parts of different wholes.

All-Around Teaching Ability

Before discussing the above results, it might be instructive to report, for each instructor, some of the correlations with an item adapted from Isaacson et al. (1963), and which served as a primary variable in their study. The item is: How would you rate your instructor in general (all-around) teaching ability? The response alternatives are: (*a*) a very poor and inadequate instructor, (*b*) a poor and inadequate instructor, (*c*) an adequate but not stimulating instructor, (*c*) a good instructor, (*3*) a very good instructor, (*f*) an outstanding and stimulating instructor, and (*g*) a very outstanding and stimulating instructor.

This item had the highest loading (.76) on the general factor of items prior to rotating to a varimax solution. The significant differences in correlations with this item, presented in Table 9, suggests that this rating has a somewhat different psychological meaning from one instructor to another. Further, with regard to this rating of all-around teaching ability, many of the differences in correlations not reaching a level of statistical significance appear divergent enough to suggest that this item would load differently on a given factor if factor analyses had been performed separately for each of the three instructor samples.

DISCUSSION AND CONCLUSIONS

By ignoring the particular context in which teacher and course ratings were obtained (i.e., using the total sample), it was possible to identify nine meaningful factors representing aspects of teacher behavior and course structure. Each factor is viewed as a positive component of an instructional situation (i.e., high factor scores are viewed as favorable instructor and course characteristics). It would seem, then, that it is possible to identify some of the general characteristics of a good (or bad) instructor and course—at

TABLE 9
CORRELATES OF THE RATING "ALL-AROUND TEACHING ABILITY" SHOWING SIGNIFICANT DIFFERENCES AMONG THE THREE INSTRUCTOR SAMPLES

Correlates	Instructor A	Instructor B	Instructor C	Significantly different correlations[a]
Teacher ratings				
Knowledge of new developments in field	.21	.53**	.32*	A versus B
Stimulates intellectual curiosity	.53**	.75**	.61**	A versus B
Sympathetic attitude toward students	.31	.33**	.03	B versus C
Sense of proportion and humor	.35*	.54**	.26*	B versus C
Course ratings				
Course well organized in meaningful sequence	.66**	.53**	.28*	A versus C
Assignments clear and reasonable	.49**	.37**	.10	A versus C
Materials suited to class level	.16	.49**	.27*	A versus B
Personality measures				
TI	−.26	.01	.23*	A versus C
Es	−.17	−.09	.33**	AB versus C
Co	−.27	−.04	.18	A versus C
RO	.11	−.30**	−.11	A versus B
Gr	.20	.22*	.57**	AB versus C

Note.—For Instructor A, $N = 39$; for Instructor B, $N = 120$; for Instructor C, $N = 68$.
[a] Differences between Z transformed correlations exceed twice the standard error of the difference. For example, with respect to Es, the correlation based on the sample for Instructor A and for Instructor B is each significantly different from the correlation based on the sample for Instructor C.
* $p < .05$.
** $p < .01$.

least as far as a course in educational psychology is concerned. A study of the personality correlates of the nine factors across the three instructors, however, does not allow this unqualified conclusion. The data of the present study consistently show that the psychological meaning of these factors varies from instructor to instructor. That is, there are significant differences in correlations between the personality and factor scores from instructor to instructor; further, different personality characteristics are correlated with a given factor for different instructors. More specifically, the type of student who tends to rate one instructor high on a given factor may be the type of student who tends to rate another instructor low on the "same" dimension. These results suggest that it is only by viewing these "abstract" factors in the light of the context in which they were obtained that their psychological meaning can be grasped. This important point can perhaps be more forcefully expressed by saying the meaning of the nine factors, considered from the perspective of the total sample, is analogous to the definitions of words found in a dictionary; the sense of these factors, as the expressive meaning of words, is determined not by the dictionary entries but by the contexts in which they occur. The instructor and the student are two aspects of the same situation; as such, the teacher's behaviors and course organization cannot be separated from the student's reactions without a considerable loss of understanding.

The most frequent personality correlates across the nine factors for Instructor A, who has a background and an interest in counseling and guidance, are Anxiety Level, Personal Integration, Response Bias, and Complexity (consistent negative correlation). A student scoring high on the first three measures and low on the fourth may be characterized as being well adjusted, as having a positive self-regard, and as disliking ambiguities and uncertainties as well as novel situations and ideas. The most frequent personality correlates for Instructor B, who has an interest in experimental and methodological procedures and issues, are Practical Outlook and Religious Orientation (consistently negative). Thus, a student who tends to evaluate ideas and things in terms of immediate utility and who has strong religious commitments will tend to evaluate Instructor B and his course favorably. Finally, the most frequent correlates for Instructor C, who has an interest in certain philosophical issues and aspects of psychology, are grades in the course and esthetic interests. The student whose ratings of Instructor C will tend to be favorable is a high achiever in the course with diverse interests in artistic matters and activities.

It is only with respect to Instructor C that the relationship between ratings and course grades supports the finding by Weaver (1960) that student ratings of teachers are related to the grades the student expected to receive in the course. Although actual course grade was used in the present study, all students knew their grades on two of the three examinations and on the paper at the time the evaluation form was completed. To speak of rater bias solely on the basis of a correlation between ratings and grades, as does Weaver, is unwarranted.

So far the emphasis has been on differences in factor correlates for the three instructors, but there are interesting similarities. If the 15 factor correlates for each instructor are ranked from high positive through zero to high negative for each of the nine dimensions, Instructors A and B are positively and significantly correlated with each other on all nine factors. Instructor C is negatively correlated with the other two on all factors, although significantly so on only four of the nine. This analysis throws into relief what the reader has probably grasped from the results already presented.

The results of the present study do not mean that the ratings are distorted by the personality characteristics of the raters and thus lack objectivity. This conclusion follows only from a view which would claim the possibility of rating from no particular point of view in no particular con-

text. Human perception requires a point of view and a context, and, in fact, they are part of the meaning of what is perceived. The data presented are consistent with this thesis. As Merleau-Ponty puts it, "The subject's intentions are immediately reflected in the perceptual field, polarizing it, or placing their seal upon it, or setting up in it, effortlessly, a wave of significance [1962, p. 131]."

By keeping the subject in perception, it is not likely that we will view him as an impartial recorder of events, and it is likely that we will understand the perceiver by noting his points of view and the correlated meanings of the situations in which we find him.

REFERENCES

CENTER FOR THE STUDY OF HIGHER EDUCATION. *Omnibus Personality Inventory: Research manual.* Berkeley, Calif.: Author, 1962.

CORCORAN, M. E. *Where does college teaching stand in the career plans of superior college seniors?* Minneapolis: Bureau of Institutional Research, University of Minnesota, 1961.

GETZELS, J. W., & JACKSON, P. W. The teacher's personality and characteristics. In N. L. Gage (Ed.), *Handbook of research on teaching.* Chicago: Rand McNally, 1963. Pp. 506–582.

ISAACSON, R. L., MCKEACHIE, W. J., & MILHOLLAND, J. E. Correlation of teacher personality variables and student ratings. *Journal of Educational Psychology,* 1963, **54**, 110–117.

KERLINGER, F. N. Attitudes toward education and perceptions of teacher characteristics: A Q study. *American Educational Research Journal,* 1966, **3**, 159–168.

MCKEACHIE, W. J. Research on teaching at the college and university level. In N. L. Gage (Ed.), *Handbook of research on teaching.* Chicago: Rand McNally, 1963. Pp. 1118–1172.

MERLEAU-PONTY, M. *Phenomenology of perception.* (Trans. by C. Smith) London: Routledge & Kegan Paul, 1962.

REMMERS, H. H. Rating methods in research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching.* Chicago: Rand McNally, 1963. Pp. 329–378.

REZLER, A. G. The influence of needs upon the student's perception of his instructor. *Journal of Educational Research,* 1965, **58**, 282–286.

VELDMAN, D. J. & PECK, R. F. Student teacher characteristics from the pupil's viewpoint. *Journal of Educational Psychology,* 1963, **54**, 346–355.

WEAVER, C. H. Instructor rating by college students. *Journal of Educational Psychology,* 1960, **51**, 21–25.

(Received February 20, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 1, 53–58

EFFECTS OF ACCELERATING BRIGHT, OLDER ELEMENTARY PUPILS—A SECOND FOLLOW-UP

HERBERT J. KLAUSMEIER
University of Wisconsin

WILLIAM L. GOODWIN
Bucknell University

AND TECKLA RONDA
Wisconsin Public Schools, Racine

Bright older children accelerated in lower elementary grades were compared with nonaccelerants toward the end of 9th grade. Ss were 22 children accelerated from Grade 2–4, 14 children accelerated from Grade 3–5, and 4 nonaccelerant groups: 27 bright younger children, 22 bright older children, 21 average-ability younger children, and 23 average-ability older children. On 6 tests of educational achievement, 9 tests of divergent thinking, and 2 psychomotor tests, both accelerant groups were equal to or higher than the other 4 groups. The nonaccelerated older bright children were higher than at least 1 of the accelerated groups on 4 tests of educational achievement, 2 tests of divergent thinking, and 2 psychomotor tests. The accelerated groups participated in school activities, advanced classes, and varsity athletics, to about the same extent as the older bright nonaccelerants.

The ideal of public education is to encourage each child to learn as well and as fast as he can, commensurate with optimal personality development. Lack of complete success in achieving this ideal is shown in the widespread attention given to academically talented children in the late 1950s and early 1960s and, more recently, to educationally disadvantaged children. This attention to various groups of children clearly indicates that instruction has not yet become sufficiently individualized to provide well for each child. Therefore, special provisions must be made on a widespread basis for groups of children. One of many possible provisions for the gifted that offers effective utilization of the resources of the school with little change in grouping procedures, organization, and the like, is acceleration whereby the student completes 12 grades of school in less than 12 calendar years.

In 1960, a random half of all the bright, older pupils who met certain specified criteria and who had just completed the second grade in Racine, Wisconsin, were accelerated to the fourth grade after a 5-week summer session. During this session, instruction related to the usual third-grade curriculum was given. Toward the end of the fourth grade, effects of this acceleration appeared to be entirely favorable (Klausmeier & Ripple, 1962). During the summer of 1961 the random half of the bright, older pupils that had not been accelerated were given the opportunity to accelerate from Grade 3 to Grade 5 after participating in a similar 5-week session, but they were not studied further. However, the first group accelerated from second to fourth grade was studied intensively again and was doing well toward the end of the fifth grade (Klausmeier, 1963). The present study considers the longer-range effects of the experience upon both groups of accelerants toward the end of the ninth grade.

METHOD

Subjects

The 129 *Ss* (54 boys and 75 girls) were distributed as shown in Table 1. The abbreviations used in Table 1 and below refer to groups as follows:

- Acc 2–4—Accelerated from second to fourth grade in 1960, currently ninth-graders.
- Acc 3–5—Accelerated from third to fifth grade in 1961, currently ninth-graders.
- 9SY—Nonaccelerated pupils of superior ability below the median age of normally progressing ninth-graders.
- 9SO—Nonaccelerated pupils of superior ability above the median age of normally progressing ninth-graders.
- 9AY—Nonaccelerated pupils of average ability below the median age of ninth-graders.
- 9AO—Nonaccelerated pupils of average ability above the median age of ninth-graders.

The students in the Acc 2–4 and Acc 3–5 groups were identified in the spring of 1960 and the others

TABLE 1

CHARACTERISTICS AND DISTRIBUTION OF SUBJECTS

Characteristic	Group					
	Acc 2–4	Acc 3–5	9SY	9SO	9AY	9AO
Mean IQ September, 1960	123.77	121.58	123.77	124.62	103.77	100.04
Mean deviation IQ May, 1966	130.9	127.6	125.0	129.9	110.1	111.4
Average age on September 1, 1965	13-5	13-5	13-11	14-3	13-11	14-4
Male subjects	9	6	11	9	8	11
Female subjects	13	8	16	13	13	12
Total	22	14	27	22	21	23

in the fall of 1960. At that time the Acc 2–4 and Acc 3–5 groups each had 16 girls and 10 boys. The same numbers of boys and girls comprised the other groups for statistical analyses although two alternate girls and boys were also identified in each group. In the present study the alternates were included in order to alleviate somewhat the natural exodus that had occurred during the six years since the study was started. As may be noted in Table 1, the smallest group is Acc 3–5 ($N = 14$). This group is smaller than Acc 2–4 partly because fewer were accelerated from Grades 3 to 5 than from Grades 2 to 4.

Also presented in Table 1 are Kuhlmann-Anderson IQ scores as of September 1960 and the mean IQ scores in deviation form as obtained from the Kuhlmann-Anderson Test, Seventh Edition, Booklet H, Personnel Press, Inc., in May 1966. The differences among the first four groups were not significant at the .05 level in 1960 nor in 1966. It is interesting to observe that the mean IQ of each group increased. One cannot determine whether the increase is solely a test artifact or whether these groups actually increased in IQ above the national standardization sample.

The mean age of each of the six groups when they began ninth grade is included in Table 1. Both Acc 2–4 and Acc 3–5 will be graduated from high school in June with a mean age of 17–2. The 9SY and 9AY groups will graduate at a mean age of 17–8, while the 9SO and 9AO groups will finish with a mean age of 18–0 and 18–1, respectively.

Instruments Used and Treatment of Data

Five types of data were assembled on the pupils near the end of Grade 9. A brief description of each instrument follows.

Educational achievement. The Tests of Academic Progress, Grade 10, Form 2, Houghton Mifflin Company, were administered to all subjects; Grade 10 of the test was given to provide an adequate ceiling for the ablest students. This test yields scores in six areas: social studies, composition, science, reading, mathematics, and literature. Raw scores were converted to standard T scores ($M = 50$, $SD = 10$) by means of the test manual; Grade 10 students were used as the norm group in conversion. Thus both accelerant groups, now in the ninth grade, are reported as having standard score means of 52 on the reading subtest since this is based on tenth-grade norms.

Ingenuity in problem solving. Form A, Ingenuity, of the Flanagan Aptitude Classification Tests (FACT), Science Research Associates, Inc., was utilized as a measure of ingenious problem solving. Although this test has a large verbal component and a rather restrictive format for gauging "ingeniousness," it nevertheless allows objective scoring. Each item right is scored 1; a maximum score is 25. High school seniors were used to norm the test, with raw scores of 14, 18, 21, and 24, falling at the 50th, 75th, 90th, and 99th percentiles, respectively.

Creative thinking abilities. Four instruments yielding eight scores were used: Alternate Uses, Form A, Expressional Fluency, Form A, Consequences, and Plot Titles, 0–1, Sheridan Supply Company. Each of the four tests was scored for fluency (number of relevant responses). Expressional Fluency and Alternate Uses were also scored for flexibility (i.e., number of relevant categories of response), while Plot Titles and Consequences were evaluated for cleverness of response. These tests were originally developed by Guilford and his associates and, although some of them are only recommended for experimental use, some available reliabilities are reported in accompanying manuals. Depending on the difficulty of the judgments to be made, a training session was held and then either two or three scorers, working independently, scored the tests. The available scores on each pupil were then averaged across scorers. The average interjudge reliability in determining each of the eight scores is reported in Table 2.

Psychomotor abilities. Three tests of psychomotor abilities were devised and administered by Grace Piskula, Consultant in Physical Education for the Unified School District, Racine, Wisconsin: Zig-Zag Run, to measure agility and large muscle coordination (the fewer the seconds required to complete the run, the higher the ability); Wall Pass, to determine eye-hand coordination and speed of reaction (the more hits of a wall with a ball in a 15-second interval, the better the coordination and reaction); and Standing Broad Jumps, to judge leg strength and ability to coordinate body parts (scores reported in inches jumped).

Participation in school activities and special

TABLE 2
AVERAGE INTERJUDGE RELIABILITES MAINTAINED BY SCORES OF TESTS OF CREATIVE THINKING ABILITIES

Test	Score	Average interjudge reliability
Expressional Fluency	Fluency	.94
	Flexibility	.95
Alternate Uses	Fluency (ideational)	.97
	Flexibility	.95
Plot Titles	Fluency (ideational)	1.00
	Cleverness	.62
Consequences	Fluency (ideational)	.98
	Cleverness	.85

$N = 2$ judges for each test except Consequences; for Consequences, $N = 3$.

programs. Each student involved in the follow-up was given a questionnaire regarding his activities, both in and out of school. The responses on these were tabulated and the percentage of the group responding yes to each item was computed. The areas of interest to investigators will become apparent in the results section, and include matters such as enrollment in "condensed" courses, honor roll lists, and participation in nonclass activities.

A 6×2 analysis of variance (groups by sex) was run on each of the measures except those under school participation. The latter are reported and discussed as percentages. Where the difference among the six groups was significant at the .05 level or beyond, a Newman-Keuls test was run to ascertain which groups were significantly higher or lower than the two accelerant groups. Differences between sets of the nonaccelerant groups (9SY, 9SO, 9AY, and 9AO) are not presented in the interest of brevity.

RESULTS

The means and standard deviations for total males and females in each group are presented in Table 3, as well as the significance level of the F ratios for groups and sex, with an indication of whether males or females were significantly higher. The mean scores on the tests of educational achievement are of interest directly because the test battery and the related norms were for tenth graders. A score of 50 is equivalent to the median score attained by the tenth graders who comprised the national standardization sample. The mean scores for both Acc groups and the 9SO group were above the tenth-grade median. A mean score of 60 is roughly equivalent to a percentile score of 85. Thus, the mean score of 62.09 made by the Acc 2–4 group in mathematics indicates very high achievement.

The differences among the means of the six groups were statistically significant on 14 of the 18 measures. The four nonsignificant differences occurred on the creative thinking battery—ideational fluency scores on Alternate Uses, Plot Titles, and Consequences—and on the Zig-Zag Run. Gener-

TABLE 3
MEANS, STANDARD DEVIATIONS, AND SIGNIFICANCE OF F RATIOS FOR 18 MEASURES

Measure	Acc 2–4		Acc 3–5		9SY		9SO		9AY		9AO		Significance of F Ratios	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	Groups	Sex
Tests of Academic Progress														
Social studies	56.6	8.4	57.3	8.2	56.1	6.9	62.0	5.8	47.0	7.9	47.9	6.6	.001	*ns*
Composition	58.3	8.2	54.4	7.7	57.0	9.1	61.4	6.5	46.0	7.3	48.6	6.3	.001	.01 F
Science	53.5	6.6	52.3	7.5	49.9	6.6	56.8	8.3	46.4	7.6	47.9	5.5	.001	.05 M
Reading	52.0	8.5	51.5	9.5	50.6	8.8	57.2	7.0	44.0	7.4	46.1	7.7	.001	*ns*
Mathematics	62.1	6.8	60.4	7.4	58.1	7.1	61.9	6.3	47.6	6.2	51.3	7.0	.001	.001 M
Literature	55.2	7.7	54.3	9.0	53.7	8.2	60.1	7.3	44.8	7.3	45.9	6.6	.001	*ns*
Flanagan Ingenuity Test	17.2	3.5	16.2	3.5	15.8	3.2	17.4	3.3	14.0	3.6	15.7	2.6	.01	*ns*
Creative Thinking Battery														
Expressional Fluency														
Fluency	8.9	3.4	8.9	3.4	7.8	2.6	10.3	2.3	6.8	3.8	7.5	4.0	.01	*ns*
Flexibility	8.1	2.8	8.7	3.4	7.6	2.5	9.9	2.5	6.4	3.5	6.4	3.3	.001	.01 F
Alternate Uses														
Fluency	25.8	6.0	22.0	6.4	25.2	8.3	26.4	8.6	23.5	7.4	20.5	7.2	*ns*	*ns*
Flexibility	24.3	4.8	21.0	6.3	23.9	7.4	25.2	7.8	20.4	7.9	19.0	7.1	.05	*ns*
Plot Titles														
Fluency	13.1	5.0	13.0	4.3	13.7	5.6	13.3	4.7	12.7	5.5	11.6	3.2	*ns*	*ns*
Cleverness	2.4	1.0	2.2	1.4	2.0	1.3	3.3	1.6	1.6	1.0	1.0	1.0	.001	*ns*
Consequences														
Fluency	50.7	16.2	55.0	19.9	57.7	23.8	57.0	17.4	49.7	18.7	50.6	13.6	*ns*	*ns*
Cleverness	21.8	7.8	22.6	11.5	21.6	7.2	25.9	8.3	17.7	8.4	14.7	4.9	.001	*ns*
Psychomotor abilities														
Zig-Zag Run	24.2	2.5	24.6	2.8	24.6	2.8	23.0	2.9	24.4	2.7	23.6	3.4	*ns*	.001 F
Wall Pass	10.4	1.9	10.7	1.6	11.4	1.9	12.2	2.6	11.3	1.6	11.2	2.7	.05	.001 M
Standing Broad Jump	67.5	10.6	67.6	10.3	69.4	10.2	74.8	13.2	69.0	7.8	73.2	14.6	.05	.001 M

ally, the 9AY and 9AO groups had the lowest mean scores on the 14 significant measures while 9SO had the highest. Especially to be noted is the fact that on 14 of the 15 tests in the cognitive domain, the 9SY group did not differ significantly from either Acc group, both of which had 1 less year of schooling and were on the average 6 months younger than 9SY. Also, the 9SO group was significantly higher than both Acc groups on only two measures and higher than one Acc group but not the other on four measures. Part of the superiority of the 9SO group may be related to a somewhat different pattern of educational experiences, to be considered more fully later. The comparison of the two Acc groups is also of interest because they had been accelerated at different times in their school careers. No difference between these two groups was significant.

The difference between the sexes, independent of the groups, was significant on seven tests. On the educational achievement measures, males were significantly higher in science and mathematics, while the situation was reversed for composition. Girls were significantly higher than boys on the flexibility score of the Expressional Fluency test. Boys performed significantly better than girls on all the psychomotor tests (girls took significantly more seconds to run a specified distance and thus performed less well than boys).

On three measures the Group × Sex interaction was significant: TAP Reading, TAP Mathematics, and the Standing Broad Jump (all $p < .05$). The girls in the Acc 2–4, 9SY, and 9AO groups scored considerably higher in reading than did boys; however, the boys were somewhat higher than the girls in the other groups—Acc 3–5, 9AY, and 9SO. No explanation for this interaction can be offered either in terms of the composition of the groups initially or their subsequent education. Both ability levels, superior and average, and both programs, accelerated and normally progressing, are equally involved. The Group × Sex interaction in TAP Mathematics was also significant. Here boys in all groups, except 9SY, had higher mean scores than girls; however, the difference between the means of boys and girls varied markedly among the five groups. The significant Sex × Group interaction for the broad jump is related to the unequal differences between the mean scores of boys and girls in the various groups inasmuch as boys were higher than girls in all groups. The large differences between boys and girls occurred in the 9SO and 9AO groups; the smaller differences occurred in the other four groups comprised of younger children. This difference is explainable in terms of physical development, namely, on measures of strength the difference between boys and girls increases with age. (The difference in running speed also increases with age but was not sufficiently large in this study to produce a significant interaction.)

Contained in Table 4 is a summary of the extent of participation by the various groups in school activities and special programs. The compressed courses were formed by presenting the normal subject matter content for 2 years in a single year. Thus, 2 years of social studies are combined in Grade 6, 2 years of mathematics in Grade 7, 2 years of science in Grade 8, and 2 years of English in Grade 9. A student completing all four courses may be taking the equivalent of tenth-grade classes as a ninth grader. As can be seen in the table, about half of each accelerant group and the 9SY group had enrolled in these courses, while three-fourths of the 9SO group had. Thus, the 9SO group had studied more tenth-grade content. The difference in participation between the accelerant and 9SO groups is particularly evident in social studies and science. Participation by the 9AO and 9AY groups in the compressed courses was very limited.

Related to the question of participation in condensed courses is that of enrollment in special summer school programs for enrichment. The percentages in Table 4 denote that half of the Acc 2–4 group, half of the 9SO group and about one-fifth of all other groups (except 9AO with about one-tenth) had attended at least one summer school session. However, the 9SO group attended about twice as many summer sessions as Acc 2–4 group and four times as

TABLE 4

SUMMARY OF PARTICIPATION IN SCHOOL ACTIVITIES AND SPECIAL PROGRAMS IN JUNIOR HIGH

Occasion	Group					
	Acc 2–4	Acc 3–5	9SY	9SO	9AY	9AO
Percentage who took compressed courses:						
Social studies	68	29	52	86	0	4
Accelerated math	73	57	48	77	0	9
Accelerated science	41	36	30	82	5	0
Accelerated English	41	50	48	55	10	17
Average	56	43	45	75	4	8
Summer school:						
Percentage who attended one or more summer sessions	50	21	19	55	19	9
Average number who attended per group member	.64	.29	.48	1.23	.19	.09
Honor roll:						
Average percentage who attained annually	23	26	17	32	5	4
Varsity teams:						
Average percentage who participated annually	15	11	24	41	13	48
Intramurals:						
Average percentage who participated annually	70	61	36	48	38	42
Activities:						
Average percentage who participated annually	21	19	22	36	17	14

many as Acc 3–5. This situation, coupled with that in the preceding paragraph, highlights the greater exposure to enriched content that the 9SO group received.

The table also indicates the percentage of each group attaining the honor roll. The percentage presented here and in the three categories below is actually an average percentage; the participation over the 3 years in junior high school has been averaged to present an annual figure. Thus, on the average, 23% of the Acc 2–4 group attained the honor roll each year, 26% of the Acc 3–5 group, etc. Although slightly more of the accelerated pupils attained the honor roll than did the 9SYs, the 9SO group placed slightly more pupils on the roll than the accelerants. Attainment of the honor roll by average pupils was negligible.

The male accelerants as a group participated in fewer varsity sports than 9SY, 9SO, and 9AO. However, the accelerants participated in more intramurals than any of the other groups, and their overall participation in sports is quite high as it was for all groups except 9SY and 9AY. Whether less participation by intellectually able boys in sports as members of a varsity team is a strength or weakness of the program is debatable at this time. The relative proportions that complete first the baccalaureate and then graduate school programs is probably more important than is current participation in athletics.

The last tabular entry reflects, again with an average annual percentage, the extent of participation in nine other activities (student council, school paper, club activities, cheerleading, science fair, talent show and other productions, orchestra, band, and chorus). As can be noted, somewhat higher participation was maintained by the 9SO group with about a third of the group being involved in each activity. All other groups had an average involvement of about one pupil of every five in each activity. From the amount of participation by the Acc 2–4 and Acc 3–5 groups, in comparison with other groups, one may infer normal sociability and social development.

DISCUSSION

At the end of the fourth and fifth grades, the effects of acceleration on the cognitive, psychomotor, and affective development of bright older children were considered generally favorable. At the end of the fifth grade, the then 5SO group was significantly higher than the Acc 2–4 group on only two of eight scores of the Metropolitan Achieve-

ment Test, none of ten tests of divergent thinking, and one of six psychomotor tests. At the end of the ninth grade, the 9SO group was significantly higher than the Acc 2–4 group on three of six educational achievement tests, one of eight tests of divergent thinking, and two of three psychomotor tests. The increased superiority of the 9SO group in educational attainment in the ninth grade, compared with the fifth, can be partially accounted for on the basis of their having taken more of the compressed courses for high-achievers during the junior high school years and also having taken more enrichment courses in summer programs. They had more opportunity to learn the subject matter included in the tenth-grade test battery used in this study. Their superiority in the physical measures probably is associated with the greater difference in physical development between the ages 14 and 13 in comparison with ages 10 and 9, the respective nearest ages of the two groups as ninth and fifth graders. In turn, the superiority in physical development was represented by the 9SO boys participating in interscholastic athletics to a greater extent.

The 9SY group is a critical comparison group at the junior high school level for they are 6 months older than the Acc 2–4, have had an additional year of schooling (at considerable additional cost), and are normally enrolled in the same classes with the Acc 2–4 and 9SO groups. The Acc 2–4 group is not significantly different from the 9SY on 17 of the 18 measures; on the mathematics test, the Acc 2–4 was significantly higher than the 9SY group. The mean score for the Acc 2–4 group is actually higher on all six educational achievement tests, on the Flanagan Ingenuity Test, and on six of the eight creative thinking tests. The Acc 2–4 group has slightly lower mean scores on the three psychomotor measures. Further, there is little difference between the Acc 2–4 group and the 9SY in a variety of school activities, including enrollment in the compressed classes and participation in social clubs or activities. The Acc 2–4 participate slightly less frequently on varsity teams but much more frequently in intramural activities. The preceding comparisons apply about equally well to the Acc 3–5 group; however, they were not studied until the ninth grade.

Not to be lost in the comparisons are the 9AO and 9AY groups. They too, have had an additional year of schooling and are considerably older than the accelerants. On five of six measures of educational attainment they are significantly lower than both Acc groups and lower than the Acc 2–4 group on the other. On none of the 15 measures in the cognitive domain is either of these groups superior to the Acc 2–4 group and only the 9AO group is superior on one of the three psychomotor measures.

These same groups will again be studied in two years and more final conclusions may be possible then. Based upon all the data collected toward the end of the ninth grade, the effects of acceleration are considered completely desirable. Some bright older children should be accelerated during the elementary school years so that they become the younger high achieving members of their classes rather than remaining the older members throughout their school life. One can predict with confidence that they will continue to be high achievers and to participate in many school activities throughout their high school years.

REFERENCES

Klausmeier, H. J. Effects of accelerating bright older elementary pupils: A follow-up. *Journal of Educational Psychology,* 1963, **54,** 165–171.

Klausmeier, H. J., & Ripple, R. E. Effects of accelerating bright older pupils from second to fourth grade. *Journal of Educational Psychology,* 1962, **53,** 93–100.

(Received February 20, 1967)

EDUCATIONAL PSYCHOLOGY
An Application of Social and Behavioral Theory

by LOUIS M. SMITH *and* BRYCE B. HUDGINS
both of Washington University, St. Louis
Knopf; 576 pages; 21 figures; 73 tables; $8.50

This is a comprehensive analysis of the principles and substantive findings of educational psychology. Although strongly emphasizing the practical problems which confront every prospective teacher, the authors also present the most significant educational theories and recent research.

SOCIETY'S CHILDREN
A Study of Ressentiment in Secondary School Education

by CARL NORDSTROM, *Brooklyn College*
EDGAR Z. FRIEDENBERG, *University of California at Davis*
and HILARY A. GOLD, *Brooklyn College*

Random House, 1967; SED 7; 224 pages; $2.45 paperbound

"A probing, daring, exciting and disquieting book which should be required reading for every teacher . . ."—Frederick E. Ellis, *Western Washington State College*

MENTAL HEALTH IN THE SCHOOLS

by THOMAS A. RINGNESS, *University of Wisconsin*
Random House, 1967; 512 pages; $6.95

A group approach in which "mental health" problems are treated as essentially problems in learning to be coped with through learning principles.

PSYCHOLOGY OF HUMAN ADJUSTMENT

by LESTER D. CROW, *Brooklyn College*
Knopf, 1967; 640 pages; $7.95

"An excellent book. Good, concise discussions, clearly presented."—Mercedes Dyer
Andrews University
An Instructor's Manual is available.

RANDOM HOUSE, Inc.
ALFRED A. KNOPF, Inc.

The College Department
501 Madison Avenue
New York 10022

1817 — Harper & Row

PYSCHOLOGICAL FOUNDATIONS OF EDUCATION

2nd Edition

Morris L. Bigge and Maurice P. Hunt

This well-documented text takes a semihistorical and comparative approach. The authors have reworked each chapter to update references, improve continuity and style, and add new material. Part I, unique in scope, treats the biology, psychology, and sociology of human nature. Part II deals with child and youth development, both physiologically and psychologically. Part III, on learning theory, presents a more penetrating discussion of this topic than can be found in competing texts at present. Part IV demonstrates the practical classroom application of theoretical and factual material in the preceding sections of the text. Instructor's Manual. *603 pages. $9.95.* ***Just published.***

PSYCHOLOGY IN THE CLASSROOM

2nd Edition

Rudolf Dreikurs

This practical manual provides the prospective and in-service teacher with the background information and methods necessary to deal effectively with behavior problems and learning deficiencies of students. Grounded in the philosophy of democracy and the socio-teleological approach of Adlerian psychology, this new edition enlarges on the significance and techniques of group approaches in the classroom, particularly the use of group discussions. *286 pages. Paper. $3.75.* ***Just published.***

THEORIES OF COUNSELING AND PSYCHOTHERAPY

C. H. Patterson

More than 125 initial adoptions within the first year of publication

This text presents summaries of fifteen major theories or points of view of counseling and psychotherapy, reviewed by the theorists discussed. Includes Alexander and French, Bordin, Ellis, Frankl, Grinker, Kelly, Dollard and Miller, Pepinsky, Phillips, Rogers, Rotter, Salter, Thorne, Williamson and Wolpe. Eleven of the fifteen theories are not available elsewhere in summary or condensed form. Extensive documentation and references. *578 pages. $9.75.*

LEARNING AND HUMAN ABILITIES: EDUCATIONAL PSYCHOLOGY
2nd Edition

Herbert J. Klausmeier and William Goodwin

Almost completely rewritten and thoroughly updated, this edition continues to emphasize the concept of emerging human abilities, thus integrating the treatment of growth and learning. Course outlines have been prepared for the instructor, and suggest a variety of combinations of all or parts of each of the 18 chapters. A Student Evaluation Guide is available to teachers. *720 pages. $8.95.*

A Student Workbook, by William Goodwin, Robert Conry, and Herbert J. Klausmeier is available for use with the text. *$3.25.*

READINGS IN LEARNING AND HUMAN ABILITIES: EDUCATIONAL PSYCHOLOGY

Richard E. Ripple

Primarily intended as a companion to LEARNING AND HUMAN ABILITIES: EDUCATIONAL PSYCHOLOGY, the READINGS are also appropriate for use with any of the current textbooks in educational psychology. Of the forty-nine readings (all unabridged), over 60% appear for the first time in a book of readings. *596 pages. Paper. $5.50.*

FOUNDATIONS OF HUMAN BEHAVIOR

Louis Kaplan

Here is a clear, concise presentation of the hard core of facts, theory and research relating to behavior adjustment, allowing the instructor to develop the implications for everyday life. *368 pages. $6.25.*

STATISTICAL CONCEPTS
A Basic Program

Jimmy Amos, Foster Lloyd Brown, and Oscar G. Mink

This brief, field-tested, constructed-response, linear program, with examples from psychology and education, presents fundamental statistics and measurement concepts in a manner especially useful to introductory courses in the behavioral sciences. *125 pages. Paper. $2.95.*

Harper & Row, Publishers • *49 E. 33d Street, N.Y.*

CURRENT ISSUES AND RESEARCH IN EDUCATION
General Editor: Harry L. Miller, *Hunter College*

This series encompasses and illuminates a wide range of new ideas, experimentation, and criticism in the rapidly expanding field of education. Each volume contains excerpts, summaries, and entire articles that survey the most recent research findings, comment on persistent issues, and evaluate continuing experiments and new ideas for the future. The editors' introductions provide the necessary historical background and underscore the relationships of the articles to general trends in the field.

1967 paper, $2.95 each volume

EDUCATION FOR THE DISADVANTAGED
Edited by Harry L. Miller

This collection of 49 articles provides an over-all view of current attitudes, research, and controversy in the area of educating the disadvantaged child. The articles stress urban and "inner city" problems in the schools, and discuss the testing dilemma, experimental projects, curriculum issues, teacher training, and new approaches to desegregation. 302 pages

THE PSYCHOLOGY OF EDUCATION
Edited by Donald H. Clark, *Hunter College*
Preface by Harry L. Miller

The 52 selections in this volume represent the most important current research in the psychology of education. The topics covered include Head Start, intelligence, emotional resources, ability grouping, and teaching the dyslexic child. The editor places special emphasis on the significance of these issues for teachers in today's schools. 288 pages

ELEMENTARY EDUCATION
Edited by Maurie Hillson, *Rutgers University*
Preface by Harry L. Miller

The 31 selections in this study discuss areas of importance in elementary education including the mathematics revolution, science, foreign language teaching, television in the classroom, and nongrading. The selections on reading include basal reading, reading at the kindergarten level, and children's literature in the classroom. 330 pages

SOCIAL FOUNDATIONS OF EDUCATION
Edited by Dorothy Westby-Gibson, *San Francisco State College*
Foreword by Harry L. Miller

This collection of 49 articles presents a comprehensive examination of the social foundations of education from the impact of social change and the pressures on the child from kindergarten through college, to the status of teachers and teacher organizations. 313 pages

from The Free Press

Fp

THE COUNSELING OF COLLEGE STUDENTS

Function, Practice, and Technique

Edited by Max Siegel, *Brooklyn College*

Foreword by Harry D. Gideonse

This manual of practical approaches to college counseling concerns itself with the everyday realities of the counselor's world. With original contributions by 18 authorities in the field, the book covers all the significant aspects of student counseling, focusing on function, practice, and technique with a depth of coverage not found in any other book on the subject.

January, 1968 — 480 pages (approx.) — $9.95

SCHOOL CHILDREN IN THE URBAN SLUM

Readings in Social Science Research

Edited by Joan I. Roberts

School Children in the Urban Slum is a collection of research studies and reviews of research in the social sciences chosen for their pertinence to educational problems encountered in urban slum schools. It is the result of work undertaken for Project TRUE (Teacher Resources for Urban Education) at Hunter College. Offering a complete survey of problems teachers face in today's urban schools, the selections are drawn from anthropology, sociology, and psychology.

1967 — 639 pages — $7.50

EDUCATION IN THE METROPOLIS

Edited by Harry L. Miller and Marjorie B. Smiley, both of *Hunter College*

Resulting from the findings of Project TRUE, this collection of readings discusses the social and economic background of the disadvantaged urban student. The selections have been chosen to provide the information and concepts necessary to understand the current problems of the urban school, and present a basis for discussing solutions to these problems. The selections are taken from the works of Michael Harrington, James Baldwin, Herbert Gans, David P. Ausubel, and others. The study is illustrated with twelve photographs by Mark Feldstein of street life in the "inner city."

1967 — 303 pages — paper, $2.95

Coming April, 1968—the companion volume to Education in the Metropolis

POLICY ISSUES IN URBAN EDUCATION

Edited by Marjorie B. Smiley and Harry L. Miller

512 pages — paper, $3.50 tent.

A New Free Press Paperback Edition—

TEACHING THE TROUBLED CHILD

By George T. Donahue and Sol Nichtern

With a New Preface by the Authors

A thoroughly tested program for the education of emotionally disturbed children within the regular school structure. ". . . a challenging and provocative approach."—*Childhood Education*

212 pages — $2.45 tent.

THE FREE PRESS

A Division of The Macmillan Company

866 Third Avenue, New York, N.Y. 10022

A sampling of recent books from Wiley

EDUCATIONAL PSYCHOLOGY IN THE CLASSROOM

Third Edition

By HENRY CLAY LINDGREN, *San Francisco State College.* A solid revision of a popular text. The Third Edition brings your students: A unique chapter on the slum child • Strong emphasis on the contributions of social psychology • A comprehensive treatment of field theory and its applications to educational practices and problems • A non-clinical consideration of mental health as it affects and is affected by problems of classroom learning.

1967 *686 pages* *$8.50*

READINGS IN EDUCATIONAL PSYCHOLOGY

By HENRY CLAY LINDGREN.

1968 *Approx. 504 pages*
Cloth: *$8.95* Paper: *$5.95*

STUDYING THE CHILD IN SCHOOL

By IRA J. GORDON, *University of Florida.* Gives a theoretical overview of child assessment in the modern school, as well as practice in using a variety of techniques for assessment.

1966 *145 pages* *$4.95* Cloth: *$4.95*
Paper: *$2.95*

CREATIVITY:

Its Educational Implications

Edited by JOHN CURTIS GOWAN, *San Fernando Valley State College;* GEORGE D. DEMOS, *California State College, Long Beach; and* E. PAUL TORRANCE, *the University of Georgia.* Readings that show how the results of creativity research can be turned to use in the school, the classroom, and the counseling office.

1967 *336 pages* Cloth: *$7.95*
Paper: *$4.95*

READINGS IN THE PSYCHOLOGY OF PARENT-CHILD RELATIONS

Edited by GENE R. MEDINNUS, *San Jose State College.*

1967 *371 pages* *$4.50*

EDUCATION AND SOCIAL CRISIS:

Perspectives on Teaching Disadvantaged Youth

Edited by EVERETT T. KEACH, Jr., WILLIAM E. GARDNER, ROBERT FULTON, *all of the University of Minnesota.*

1967 *413 pages* Cloth: *$7.95* Paper: *$4.95*

JOHN WILEY & SONS, Inc.

PARENTS LEARN THROUGH DISCUSSION

Principles and Practices of Parent Group Education

By ALINE B. AUERBACH, *formerly of The Child Study Association of America.* This source book on the philosophy, goals and techniques of the group education method for parents is written by an eminent authority in parent education. Drawing upon her years of experience in developing and directing parent group education programs for The Child Study Association, she presents a detailed, practical guide for setting up and conducting discussion groups for parents and expectant parents. Attention is also given to groups for parents whose children are physically or emotionally handicapped, unwed mothers, and adoptive parents and others.

Considerable stress is given to the group experience of parents from low socioeconomic and educational backgrounds. Much first-hand evidence shows how these and other parents make use of educational group programs when they are clearly geared to their immediate situations and present needs.

Although the group methods described here pertain to work with parents, the principles and techniques are applicable to many group programs. *1968* Approx. *384 pages $7.95*

A HISTORY OF GENETIC PSYCHOLOGY

The First Science of Human Development

Edited by ROBERT E. GRINDER, *University of Wisconsin.* Traces concepts of fertilization, heredity, and growth from the early Greeks to G. Stanley Hall—the father of the child study movement in America. *1967 247 pages $8.95*

STUDIES IN COGNITIVE GROWTH

By JEROME S. BRUNER, *Harvard University; with* ROSE R. OLVER, *Amherst College*; PATRICIA M. GREENFIELD, *et al.* The first major theoretical assessment of the process of cognitive development in children since the pioneering work of Jean Piaget and his colleagues at the University of Geneva. *1966 343 pages $7.95*

CLASSROOM GROUPING FOR TEACHABILITY

By HERBERT A. THELEN, *University of Chicago.* The result of a three-year research investigation, this important new book presents the rationale and procedure for "compatibility" grouping and cites evidence of its validity. *1967 274 pages $7.50*

605 Third Avenue, New York, N. Y. 10016

Looking for better textbooks?

Educational Psychology Third Edition

By Glenn Myers Blair, R. Stewart Jones, and Ray H. Simpson,

all, The University of Illinois

This successful text, now in its Third Edition, deals with the problems considered most important by teachers and school psychologists. It includes new material on the psychology of the teacher and teacher self-appraisal. The book is unusual for its in-depth treatment of psychology of adolescence, the social psychology of teaching, and diagnostic and remedial procedures. Principles of learning are constantly supported by experimental evidence. Psychological theories are illustrated by actual classroom examples.

Intended for use in undergraduate courses, the book presents a comprehensive, thoroughly documented and highly practical approach to problems confronting teachers and psychologists. While it is of value to both, it is addressed particularly to the teacher in his role of responsibility for preparing the child to enter the complex, technological modern world.

Clearly and imaginatively written, this book gives the reader practical, up-to-date help in engineering the learning process in a manner consonant with sound psychological theory. A Teachers' Manual is available.

1968, 704 pages, $8.50

Readings in Educational Psychology Second Edition

Edited by Victor H. Noll and Rachel P. Noll, both, Michigan State University

This thoroughly revised collection of readings provides the student in educational psychology with a source of significant and relevant literature in the field. Twenty-eight of the articles are new and reflect a changing emphasis in educational psychology. Especially written for this anthology are articles by Robert C. Craig, Robert L. Ebel, Ruth Strang, Elizabeth M. Drews, and William Gnagey.

1968, 464 pages, paper, $4.25

Write to the Faculty Service Desk for examination copies.

THE MACMILLAN COMPANY 866 Third Avenue, New York, N.Y. 10022

education is a happening, and here's what's happening in education

PSYCHOLOGY FOR EFFECTIVE TEACHING
Second Edition

By George J. Mouly, University of Miami

Revised and updated to highlight recent developments of major psychological significance, the Second Edition of this text emphasizes (1) the recent awareness of the crucial role of environmental influences in the development of one's potentialities, (2) the new orientation toward the stimulation theories of motivation, and (3) the current interest in creativity and the multi-dimensional nature of the human intellect.

January 1968 | 656 pp. | $8.95

EDUCATIONAL PSYCHOLOGY
A Cognitive View

By David P. Ausubel, University of Toronto

The basic premise of this book is that educational psychology is primarily concerned with the nature, conditions, outcomes, and evaluation of classroom learning. Unlike most other books in the field, it does not treat educational psychology as a loose amalgam of learning theory, developmental psychology, mental hygiene, and educational or psychological measurement, or as a simplified version of general psychology.

May 1968 | 640 pp. | $8.95 (tent.)

THE PROCESS OF SCHOOLING
A Psychological Examination

By J. M. Stephens, University of British Columbia

Professor Stephens explores the phenomena of learning as a process that somehow continues in spite of innovations and changes. This point of view represents the author's belief that education is a natural process which teachers must try to abet as well as they can—a position which is intended to provoke some different, new ways of looking at the educational work teachers are trying to do.

1967 | 176 pp. | $3.95 paper

Holt, Rinehart and Winston, Inc.

383 Madison Avenue, New York, New York 10017

McGraw-Hill Books

CLASSICAL PSYCHO-PHYSICS AND SCALING

By Sidney A. Manning, University of Maryland; and Edward Rosenstock, Pennsylvania State University.

176 pages, $2.50

This is a semi-programmed text designed to teach psychophysics and scaling as part of a general introductory course, or as an adjunct to courses in Perception, Experimental, or Physiological Psychology.

READINGS IN PSYCHOLOGICAL FOUNDATIONS OF EDUCATION

By Walter H. MacGinitie and Samuel Ball, Teachers College, Columbia University

Available Spring, 1968

The authors have selected articles that are understandable to college students with little or no background in psychology or statistics. The book is especially effective for use in a one-semester course.

CLINICAL TEACHING

By Robert M. Smith, Pennsylvania State University

Available Spring, 1968

In this text, the author considers the educational problems of the retarded child from the pre-school through the post-school periods. Suggestions are made for working with parents of the retarded within the structure of the school program.

Examination copies upon request

McGraw-Hill Book Company

330 West 42nd Street,
New York, New York 10036

ACC

Works of major interest

THE TECHNOLOGY OF TEACHING

B. F. SKINNER, *Harvard University.* Four papers already on the way to becoming classics, together with a number of new papers appearing now for the first time, here embody B. F. Skinner's long-awaited critique of current educational philosophies. *271 pp., $5.50; paper, $2.95*

EDUCATIONAL PSYCHOLOGY

JANICE GIBSON, *University of Pittsburgh.* This graded program provides both students and in-service teachers at all levels with an overview of those areas of psychology which are relevant to education. It contains all of the problems studied by the educational psychologist while it reflects the point of view of a psychologist experienced both in research methods of psychology and in applications to classroom problems. *March 1968, 416 pp., 1375 frames, $4.95 (tent.)*

Appleton-Century-Crofts

Div. of Meredith Corporation

440 PARK AVE. S., N. Y. 10016

New from BASIC BOOKS

PSYCHOLOGICAL CONSULTATION IN THE SCHOOLS

A Catalyst for Learning

Ruth G. Newman

Through personal observation, Dr. Newman has developed realistic guidelines for setting up a continual, on-the-spot psychological consultation program at every school level. Since the problems in the schools studied by Dr. Newman—with their population of both the middle class and the Negro poor—are likely to be found in any urban community, the method proposed in this book can be adapted to initiate or strengthen consultation programs in any metropolitan complex. "I know of few publications, theoretical in basic intent, in which the flavor of actual child life in school is so genuinely preserved and which give the reader such an intense feeling of 'having been there.' " —Fritz Redl, Wayne State University $6.95

THE DISADVANTAGED CHILD

Studies of the Social Environment and the Learning Process

Martin Deutsch and Associates

Offering a wealth of empirical data and a rich store of theoretical insights, Martin Deutsch (often referred to as the "father" of Head Start) and his colleagues explore the psychological and social environments of disadvantaged children, and examine their effects on normal development and their implications for education. ". . . a deep reservoir of ideas . . . should be read by all American educators concerned about creating a better society based on better schools, with equality of opportunity for every child." —Marian Wozencraft, State University of New York $10.00

BASIC BOOKS, INC., PUBLISHERS

404 Park Avenue South, New York, N.Y. 10016

A new revision

of major significance . . .

APA PUBLICATION MANUAL

1967 Revision

The APA PUBLICATION MANUAL has undergone a complete revision for the first time in 10 years. Although the revised edition prescribes the same basic format for manuscripts as the previous edition, major changes and additions were made in the text in an effort to clarify and simplify the instructions on how manuscripts should be prepared.

The APA PUBLICATION MANUAL is intended primarily as a guide for authors and editors in the preparation of material to be published in APA journals. However, in preparing the original edition it was recognized that the MANUAL could, and should, be written to accommodate a wider range of manuscripts. The revised PUBLICATION MANUAL has retained that original concept. $1.50

Order from:

AMERICAN PSYCHOLOGICAL ASSOCIATION

1200 Seventeenth Street, N. W.

Washington, D. C. 20036

Journals of the

AMERICAN PSYCHOLOGICAL ASSOCIATION

Subscription Rates for 1968

American Psychologist. Official papers of the Association and articles on psychology. Monthly. The 1968 volume is Vol. 23. Subscription: $10.00 (foreign $10.50). Single copy $1.00

Contemporary Psychology. Critical reviews of books, films, and research material in the field of psychology. Monthly. The 1968 volume is Vol. 13. Subscription: $10.00 (foreign $10.50). Single copy $1.00.

Journal of Abnormal Psychology. Theoretical and research contributions in behavior pathology. Bimonthly; first issue appears in February. The 1968 volume is Vol. 73. Subscription: $10.00 (foreign $10.50). Single copy $2.00. (Continues volume sequence of the former Journal of Abnormal and Social Psychology).

Journal of Applied Psychology. Applications of psychology in business and industry. Bimonthly; first issue appears in February. In addition to the regular journal issues it is expected that subscribers will receive two Monograph supplements annually. The 1968 volume is Vol. 52. Subscription: $10.00 (foreign $10.50). Single copy $2.00.

Journal of Comparative and Physiological Psychology. Original contributions in the field of comparative and physiological psychology. Bimonthly; first issue appears in February. Two volumes per year, three issues in each volume. In addition to the regular journal issues it is expected that subscribers will receive two Monograph supplements annually. The 1968 volumes are Vols. 65 and 66. Subscription: $20.00 for 2 vols. (foreign $20.50). Single copy $4.00.

Journal of Counseling Psychology. A primary publication medium for research in counseling theory and practice. Bimonthly; first issue appears in January. The 1968 volume is Vol. 15. Subscription: $10.00 (foreign $10.50). Single copy $2.00. (Vols. 1–13 published by Journal of Counseling Psychology, Inc.).

Journal of Consulting Psychology. Research in clinical psychology: psychological diagnoses, psychotherapy, personality, psychopathology. Bimonthly; first issue appears in February. In addition to the regular journal issues it is expected that subscribers will receive two Monograph supplements annually. The 1968 volume is Vol. 32. Subscription: $10.00 (foreign $10.50). Single copy $2.00.

Journal of Educational Psychology. Studies of learning and teaching: measurement of psychological development, psychology of school subjects, methods of instruction, school adjustment. Bimonthly; first issue appears in February. The 1968 volume is Vol. 59. Subscription: $10.00 (foreign $10.50). Single copy $2.00.

Journal of Experimental Psychology. Original contributions of an experimental character. Monthly. Three volumes per year, four issues in each volume. In addition it is expected that subscribers will receive approximately six Monograph supplements annually. The 1968 volumes are Vols. 76, 77, and 78. Subscription $30.00 for 3 vols. (foreign $30.50). Single copy $3.00.

Journal of Personality and Social Psychology. Theoretical and research papers on personality dynamics, group process, and the psychological aspects of social structure. Monthly. Three volumes per year, four issues in each volume. In addition it is expected that subscribers will receive approximately six Monograph supplements annually. The 1968 volumes are Vols. 8, 9 and 10. Subscription: $30.00 for 3 vols. (foreign $30.50). Single copy $3.00.

Psychological Abstracts. Noncritical abstracts of the world's literature in psychology and related subjects. Monthly. Approximately 17,000 abstracts published annually. The 1968 volume is Vol. 42. Subscription: $30.00 (foreign $30.50). Single copy $3.00.

Psychological Bulletin. Evaluative reviews of research literature, discussions of research methodology in psychology. Monthly. Two volumes per year, six issues in each volume. The 1968 volumes are Vols. 69 and 70. Subscription: $20.00 for 2 vols. (foreign $20.50). Single copy $2.00.

Psychological Review. Original contributions of a theoretical nature. Bimonthly; first issue appears in January. The 1968 is Vol. 75. Subscription: $10.00 (foreign $10.50). Single copy $2.00.

(**Psychological Monographs.** Discontinued as a separate subscription. Monographs distributed as supplements to appropriate subject area journal.)

Cover page, table of contents, and index are contained in the last issue of the volume for each journal.

NOTE: There are no quantity, cash, library, or other special discounts. There is no club arrangement.

★ Subscriptions are available on a January through December basis only ★

Payment must be made in U.S. funds.
Send subscription orders to:

AMERICAN PSYCHOLOGICAL ASSOCIATION

Subscription Department
1200 Seventeenth Street, N.W.
Washington, D.C. 20036

JOURNAL OF
EDUCATIONAL PSYCHOLOGY
MONOGRAPH SUPPLEMENT

EDUC. & PSYCH. LIBRARY

Vol. 59, No. 1, Part 2 February, 1968

DISCOVERY AND EXPOSITORY TASK PRESENTATION IN ELEMENTARY MATHEMATICS[1]

BLAINE R. WORTHEN[2]
University of Utah

This study compared 2 methods of task presentation which differed primarily in terms of sequence characteristics. 432 5th- and 6th-grade Ss were presented with 6 weeks of instruction in elementary mathematics through textlike experimental sequences introduced by classroom teachers trained in both discovery (Treatment D) and expository (Treatment E) sequencing. Analysis revealed equal teacher adherence to prescribed teaching behavior models in each treatment. Treatment E proved superior on initial learning, while Treatment D was superior on retention and transfer of heuristics. Treatment D was slightly superior to Treatment E on a test of transfer. Given equal time on the learning task, Ss in Treatment D proved superior to Ss in Treatment E on a majority of intertreatment comparisons.

The past decade has been marked by a continuing controversy over the relative efficacy of differential methods of task presentation which have been loosely referred to as "discovery" and "expository" methods. Adherents of the discovery method argue strongly for its superiority as a method of teaching and claim that "discovery learning" enhances retention and transfer of concepts as well as pupil motivation (Beberman, 1958; Bruner, 1960). Critics of the discovery method discount it as pedagogically impractical and argue that it offers little to the learner that cannot be offered equally well by good expository teaching (Ausubel, 1961, 1964).

Apparent support for each of these positions can be found in the research literature. For example, Swenson (1949) and Ray (1961) conducted comparative studies in which "discovery" methods produced significantly better results on retention and transfer measures than did "expository" methods. But Craig (1956), Kittell (1957), and Kersh (1962) found just the opposite to be true. Investigations by Hendrix (1947) and Gagné and Brown (1961) support discovery as superior to exposition in terms of transfer of learning. But studies by Craig (1953) and Corman (1957) favor expository techniques on transfer tests.

Perhaps the greatest factor which contributes to such equivocal research evidence is the differing specification among researchers as to what they mean by such terms as "discovery," "guided-discovery,"

[1] This investigation was supported by the Cooperative Research Program of the Office of Education, United States Department of Health, Education and Welfare—Project 2277 and constitutes part of the final report of that project (Della-Piana, Eldredge, & Worthen, 1965). The data collected for this investigation also served as an essential portion of a master's thesis (Worthen, 1965) submitted to the Department of Education, University of Utah. For a more complete review of related research and a detailed description of all methods, analyses, instrument validation, results, etc., the reader is referred to either of the above sources, or the American Documentation Institute. The curriculum materials used with both the discovery method and the expository method, instruments used for evaluating the outcome of the experiment, and several tables of additional data have been deposited with the American Documentation Institute. Order Document Number 9633 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington, D. C. 20540. Remit in advance $26.25 for photocopies or $7.50 for microfilm and make checks payable to: Chief, Photoduplication Service, Library of Congress.

[2] Now at the Ohio State University.

© 1968 by the American Psychological Association, Inc.

and "exposition." Since these terms have not yet been reduced to generally accepted operational definitions, it is highly probable that researchers working in what is nominally the same domain are not actually investigating the same phenomena at all. As Wittrock (1966) points out, in an excellent review and analysis of the literature on learning by discovery, semantic inconsistency in labeling differing treatments with the same name has been a major factor in precluding any important conclusions about learning or teaching from being drawn from such research. Many investigators have been primarily concerned with the amount and type of external guidance to which the learner is subjected. Others have been concerned chiefly with the role of verbalization in discovery-expository processes. Still other researchers have focused on feedback mechanisms or on rate of presentation. Such wide divergence in the variables controlled in various studies has led to investigation of widely differing facets of the discovery and expository processes and a consequent noncomparability of the results (Della-Piana, Eldredge, & Worthen, 1965).

Previous studies have been almost wholly exploratory in nature, however, and it is not surprising that direct comparisons among the results are impossible at present. Much more exploration is necessary before it will be possible to discern which, if any, comparisons are legitimate. The relevant teaching-learning variables must be more fully identified and interrelated and systematic research conducted on interactions of method with differing types of teachers, pupils, and subject matter. Until this is done, no unequivocal pattern can be hoped to emerge from discovery-expository research.

While some of the relevant variables have been extensively explored, others have received relatively little attention. One variable which has been largely ignored in previous "discovery" studies is that of the sequence characteristics of the learning tasks. It could be argued that the type or amount of external guidance or verbalization is no more important in concept formation than the sequencing of such guidance or verbalization. Certainly this aspect of discovery teaching deserves investigation in its own right. One major purpose of the present study was to take a first step in such investigation by describing and comparing a discovery method and an expository method which differed in terms of the sequence characteristics of the presentations.

Most "discovery" studies have been conducted in a laboratory setting and consequently have dealt with small time samples, small numbers of subjects (*Ss*), and very discrete and often manipulative learning tasks. One might argue that such sampling of time, *Ss*, and tasks is so restrictive and limited in scope that any attempt to generalize the results to classroom learning or instruction would be subject to serious question. It would seem that the results of a carefully-controlled classroom experiment where both time sample and learning task are representative of typical school behavior and curriculum could be generalized to classroom practice with more confidence than could the results of the typical short-term laboratory experiment.[3] A second major purpose of the present study was to compare the two instructional methods in a naturalistic setting where the learning tasks and time sample approximated normal classroom conditions. Accordingly, certain concepts in elementary mathematics were selected as content for the two differing instructional sequences. These sequences were presented to the *Ss* through textlike instructional programs introduced into the classroom by teachers trained in both experimental methods.

The criteria used to measure the outcomes of instruction included the following: tests of initial learning, retention, and transfer of the selected mathematical concepts; tests for transfer of heuristics; and measures of attitude toward the subject content. A complete listing and brief de-

[3] The difficulty of controlling research in a naturalistic classroom setting has been documented (Bellack, Davitz, Kliebard, & Hyman, 1963, pp. 165–168; McDonald, 1964, p. 542) and is acknowledged by the investigator. It would seem, however, that difficulty should not prevent attempts to find productive ways to utilize the classroom as a research setting.

scription of these criterion measures appears in the section "Tests and Measures."

Hypotheses

Although no concrete conclusions can be drawn from the hodgepodge of research which has been conducted on discovery learning, some plausible inferences can be made.

One such inference is that short-term studies should tend to favor expository instruction while long-term studies should favor discovery learning. This inference is based on the assumption that pupils have typically been trained to learn from expository instruction and, hence, need a relatively longer period of time in which to learn to utilize the techniques necessary in discovery learning. Based on this assumption, and in view of the length of the present study, it was hypothesized that the discovery method used in this study (Treatment D) would produce superior results to the expository method (Treatment E) on tests of initial learning, retention, and transfer of the selected mathematical concepts.

An additional hypothesis was based partly on work by Kersh (1958, 1962), which suggested that any advantages of discovery learning could best be explained in terms of pupil motivation, and partly on the experimenter's experience with pupils in the initial tryout of the instructional materials. During this tryout, pupils who were exposed to the Treatment D instructional sequences made considerably more unsolicited statements indicating that they liked the "new materials" than did pupils exposed to the Treatment E instructional sequences. It was hypothesized that Treatment D would produce superior results to Treatment E on measures of attitude toward the subject content.

A final hypothesis was based on a logical extension of interpretations by Suchman (1962) and Della-Piana (1957) relating to searching set and its effect upon conceptual data processing. It was hypothesized that Treatment D would produce superior results to Treatment E on tests for transfer of heuristics.

METHOD

Treatments

Brief definitions of the experimental methods appear below.

Discovery method (Treatment D). Treatment D is defined as a method in which verbalization of each concept or generalization is delayed until the end of the instructional sequence by which the concept or generalization is to be taught. The pupil is presented with an ordered, structured series of examples of a generalization. The sequence of presentation maximizes the possibility of the pupil formulating awareness of the generalization more readily than if the examples are randomly presented. No explanation of the examples is given, nor is there any hint that there is an underlying principle to be discovered. The pupil, merely instructed to solve the problems, is expected to acquire the mathematical concept, principle, or generalization through an inference of his own.

Expository method (Treatment E). Treatment E is defined as a method in which the verbalization of the required concept or generalization is the initial step in the instructional sequence by which the concept or generalization is to be taught. The mathematical principle is presented to the pupil and explained verbally and symbolically. The pupil works with examples of the principle or generalization only after the initial verbal and symbolic presentation. Particular attention is given to insure that practice is made meaningful by continual stress being placed upon the relation of the example to the generalization and upon "why" the generalization operates as it does. This is to minimize rote memorization of the principle by the pupil.

Subjects

The *Ss* were 538 fifth- and sixth-grade pupils in the Salt Lake City School District. The experimental sample was comprised of 432 of these pupils who were equally divided among 16 classes.[4] These classes were equally divided among eight elementary schools which were nominated by district central office personnel as representative of the elementary schools in the district in terms of socioeconomic and geographical characteristics.

The teachers were selected on the basis of the following criteria: (*a*) mathematical and general

[4] A control group, comprised of 106 pupils in 3 sixth-grade classes, received both the pre- and posttests but received no special instruction during the intervening 6-week period. This group was included in the study in order to provide normal baseline data against which to assess effects of the two experimental treatments. Results of the intertreatment comparisons between the experimental groups and the control group appear in detail in previous reports of this research but are omitted here in the interest of brevity. It should be noted, however, that the results of these comparisons support the findings and conclusions reported herein.

teaching competence as judged by supervisors, (*b*) minimum of 3 years of teaching experience, and (*c*) willingness to participate in this research project. The selection of the teachers determined the selection of the sample; *S*s used in this study were pupils in established classes of the selected teachers.

Experimental Design and Controls

In each of eight schools, two classes were taught arithmetic by the same teacher, one class by Treatment D and one class by Treatment E. This was done in an attempt to control the dimensions of teacher personality and other teacher characteristics. Seven of the teachers taught two sixth-grade classes each while the eighth teacher taught two fifth-grade classes. Seven of the eight teachers taught their own homeroom class as one of the experimental groups. In an attempt to control possible differential in pupil-teacher interaction between homeroom and nonhomeroom classes, the number of homerooms receiving each treatment was balanced as nearly as possible. The assignment procedures also balanced as nearly as possible the number of classes receiving each treatment during any particular segment of the school day. Although there was no reason to believe that the selection and assignment procedures would bias the sample, a preliminary inspection of the mean values for each treatment group was conducted on several pretreatment measures including IQ, arithmetic computation skill, arithmetic problem-solving ability, prior knowledge of the selected mathematical concepts, prior attitude toward arithmetic, and pupil perception of teaching behavior. The only significant differences found between the experimental groups were on the attitude measures. The *S*s in Treatment E entered the experimental period with significantly more favorable attitudes toward arithmetic than *S*s in Treatment D.

The major nonexperimental variables controlled in this study are presented below.

1. The *S*s in Treatments D and E received the same length of time to work on the learning tasks.
2. The amount of verbalization in the teachers' oral presentation and in the written instructional materials was held constant in both treatments. Verbalization of the mathematical generalizations varied in sequence between the two treatments but was present in both.
3. Three techniques were used in this study in an attempt to assess the extent to which the teachers taught by the specified methods. These techniques (utilizing instruments described later) included the following: (*a*) rating by observer-raters of a 10% sample of the total teaching behavior of each teacher in each treatment; (*b*) rating of a 10% sample of total teaching behavior of each teacher in each treatment from lessons recorded on audio-tape; and (*c*) rating by pupils of teaching behavior on the discovery-expository dimension.
4. All of the procedures and methods utilized in this study were selected in an attempt to minimize or negate any differential "Hawthorne effect" between the two experimental groups. Directions given to both treatment groups were as nearly identical as possible. All procedures used in one treatment were used in the other, and both groups were made to feel that they held shared status with relation to the "new math project."
5. An attempt was made to equalize the pre-experimental mathematical experiences of all *S*s in Treatments D and E by presentation, during a 2-month period immediately preceding the pretests, of an unit which included both specific and general mathematical concepts judged to provide necessary background for the experimental materials. In addition, confounding of the experimental results by nonexperimental arithmetic experiences was minimized by a request that no homework or out-of-school arithmetic assignments be given. District personnel complied with this request and also elicited parental cooperation.

The experimental period, 40 minutes daily for each treatment, consisted of 3 days of pretests, 6 weeks of instruction, and 5 days of posttests.

Training Program for Teachers and Raters

All raters and teachers attended a training class which met from 2 to 6 hours weekly for 20 weeks, 13 weeks prior to and 7 weeks during the experiment. Training was given in: (*a*) general mathematical concepts necessary as background; (*b*) all mathematical concepts used in the instructional materials and criterion measures; (*c*) use of the two specific methods of instruction; and (*d*) procedures for administering and scoring the various tests, scales, and questionnaires.

Instructional Materials

The instructional materials presented several mathematical concepts selected on the basis of suitability for both discovery and expository teaching and probable unfamiliarity to *S*s at the inception of the study. The concepts selected were the following: (*a*) notation, addition, and multiplication of integers (positive, negative, and zero); (*b*) the distributive principle of multiplication over addition; and (*c*) exponential notation and multiplication and division of numbers expressed in exponential notation.

The materials were equated in terms of the mathematical concepts, diagrams of physical models, number and type of examples, and degree of verbal presentation used in each treatment. The two sets of materials differed primarily in terms of sequence characteristics.

Instructional Procedures

The instructional procedures in each treatment were largely determined by the requirement that the teachers follow the structural sequences of the instructional materials. However, some aspects of teaching behavior were judged to be independent of the instructional materials but still influential in maintaining the essential sequence differences

between the two treatments. The characteristics of teaching behavior which seemed most operative in this regard include the following: (*a*) interjection of teacher knowledge, (*b*) introduction of generalization, (*c*) method of answering questions, (*d*) control of pupil interaction, and (*e*) method of eliminating false concepts. Model "discovery" teaching behavior and model "expository" teaching behavior on each of these five characteristics was specified and a paradigm of teaching techniques for each characteristic was established in each treatment. A brief summary of model teaching behavior for each treatment on each of the five characteristics of teaching behavior is given below.

1. *Interjection of teacher knowledge—Treatment E.* The teacher acts as the primary source of knowledge concerning arithmetic. The students may depend on the teacher when they cannot work a problem. The teacher always indicates that he will show the student how to work the problem correctly. He is never doubtful or uncertain. The teacher checks the answers of the students and shows them how the correct answer is obtained by use of the principle involved. When an incorrect answer is given, the teacher recognizes it and immediately asks the student if he is certain that his answer is correct. This gives the student an opportunity to correct his own mistake. If the student is unable to do so, the teacher asks for someone else in the class to respond. Care is taken, however, to avoid any negative evaluation of the student's response.

Treatment D. The teacher does not act as the primary source of knowledge concerning arithmetic, but seems to depend upon the students to help him work the problems. He sometimes reflects an uncertain feeling about the precise way to solve a particular problem. When given an answer, whether correct or incorrect, he checks it by the long method as if he is not aware of the principle which would allow allow for solution by a "shortcut." When a student gives an incorrect answer, the teacher continues on to the next problem as if unaware that the answer is incorrect. When a student points it out, the teacher acts surprised. If the students fail to notice the mistake, the teacher goes back a short time later, as if he has just noticed it, and questions the correctness of the earlier response. The student who gave the response is allowed an opportunity to correct it. If he is unable to do so, other members of the class are asked to respond.

2. *Introduction of generalization—Treatment E.* The teacher gives the generalization (rule) before the students are given any examples. All examples are then related back to the rule for solution.

Treatment D. The teacher delays the verbalization of the generalization until all, or nearly all of the students have made the discovery. He is careful to give no hint that there is a "shortcut" to working the problems. He also takes care to avoid the use of vocabulary terms related to the generalization.

3. *Method of answering questions—Treatment E.* The teacher answers questions by reiterating and explaining the rule and relating it to the question. The teacher then gives examples which will further clarify the way the rule is used in the solution of that type of problem.

Treatment D. The teacher answers questions by referring to the model or the computational sequence which the student has used. If a student is still confused, the teacher takes him back to the model and goes through it carefully. The teacher may make use of sequenced examples as a clue, but no verbal hint of the rule is given.

4. *Control of pupil interaction—Treatment E.* The teacher allows the students to share ideas about arithmetic. He encourages them to help each other if they can show the other person how to apply the rule. Often he allows them to check an answer with their neighbor but insists that they do their own work.

Treatment D. The teacher always prevents the students from sharing ideas about arithmetic. The teacher insists that no student does anything which may inhibit another child's discovery by giving the rule to him prematurely. If a student does verbalize the rule, the teacher expresses displeasure.

5. *Method of eliminating false concepts—Treatment E.* The teacher warns the students of common errors made in applying the principle. He points out specifically the types of problems the students are likely to make errors on and then gives examples of each kind of error.

Treatment D. The teacher includes "trap questions" and gives no verbal warning of any type. The teacher purposely leads the child, through sequencing of examples, into overgeneralizing the rule. If the problem is missed, the teacher waits until the error is mentioned, then checks to make certain it is wrong. He says nothing about the rule or why the answer was incorrect.

Adherence to the model techniques of teaching specified for each treatment and to the sequence of presentation determined by the instructional materials was assessed by observer and pupil rating scales (described hereafter). Scores on these scales were used as an index of teacher fidelity in the presentation of the treatment.

Because of the wide range of ability among and within classes, teachers were allowed to vary their rate of instruction in order to fit the needs of their particular class. Although the total time consumed by each treatment was held equal, how far each class progressed in the instructional materials was dictated by the rate of instruction. Teachers were required to cover each concept and principle carefully, using the prescribed teaching techniques, following the sequence dictated by the materials, and making every attempt to make both treatments equally meaningful. In order to insure adequate presentation of the concepts to both treatment groups, the criterion was established that a minimum of 85% of each class must attain a specified minimum level of understanding of each concept before the teacher was allowed to proceed to the next concept. Criterion items for each concept were

TABLE 1

RELIABILITY ESTIMATES FOR SEVEN INSTRUMENTS

Instrument	Reliability			
	Test-retest		Parallel form	
	Consec. days	6-week interval	Consec. days	6-week interval
Semantic Differential Attitude Scale	.87	.44		
Statement Attitude Scale	.78	.75		
Pupil Perception of Teaching Behavior	.92	.92		
Concept Knowledge Test	.70[a]	.73[a]		
Concept Retention Test	.75[a]		.75[a]	.69[a]
Concept Transfer Test	.82			
Negative Concept Transfer Test	.64			

Note.—For the consecutive days, N = 57, for the 6-week interval, N = 106.

[a] Reliability coefficients on these instruments are somewhat difficult to interpret. Because the content measured by these instruments was generally unique to fifth- and sixth-grade pupils, only those pupils used as *Ss* in the experimental treatments and thus exposed to the content were able to score well consistently. Pupils in the reliability reference groups achieved very low scores and exhibited little variation. In view of the direct relationship between the magnitude of a reliability coefficient and the variation of the sample on which it is based, the reliability estimates reported for these instruments were judged to be completely acceptable.

built into the materials to enable this test to be made. The application of this criterion resulted in some variation among schools in the amount of the instructional materials completed during the experimental period. There was virtually no variation between treatments, however, when summed across schools. In seven of the eight schools, the number of instructional units completed by the two treatment groups was equal. In the eighth school, the Treatment E class completed two pages more of the instructional materials than the Treatment D group in the same period of time. This difference was discounted by the experimenter as inconsequential.

Tests and Measures

Ten instruments were developed for this study, nine of which were administered to all *Ss* while the tenth was used to rate teacher behavior. Reliability coefficients for seven of these instruments appear in Table 1. Reliability of the other three instruments is discussed below in connection with the description of those instruments. Additional reliability information and material related to validity of the instruments can be found in the original report of this research (Worthen, 1965).

Prior knowledge of the selected mathematical concepts was measured by a test (Concept Knowledge Test) administered to both treatment groups in the pretest series. Initial learning was measured by the four subsections of this same test administered at the completion of the corresponding subsection of the instructional materials. A parallel form of this test (Concept Retention Test) was administered to both treatment groups 5 weeks and 11 weeks after the first administration in order to measure retention. Both instruments were also administered to the reliability reference groups. Scores on these two forms were correlated and the coefficients, reported in Table 1, bear out the claim of parallelism. The correlations between forms are approximately equal to the reliability coefficients reported for either form.

A positive transfer test (Concept Transfer Test) was administered to both treatment groups in the posttest series and was used to evaluate *Ss'* ability to recognize and apply mathematical principles in situations unlike those in which they were originally presented. A negative transfer test (Negative Concept Transfer Test) was added in order to assess *Ss'* tendency to overgeneralize the principles to inappropriate situations.

Transfer of heuristics was measured by two tests. The first of these (Written Heuristic Transfer) was designed to assess the effects of the two treatments on *Ss'* ability to discover a mathematical principle on a "paper and pencil" task comprised of a series of written problems, each of which could be solved easily if *S* discovered the "shortcut." The second test (Oral Heuristic Transfer) consisted of a sequence of problems presented orally by the teacher, each of which could be solved easily if *S* discovered the "shortcut." On this test, final criterion behavior was determined by performance on a six-problem paper and pencil exercise. The major difference between the tests was in the mode of presentation (oral or written) of the initial problem sequence. Both of these tests were administered in the posttest series to *Ss* in both treatments. Because of the nature of these two instruments, it was impossible to obtain a reliability coefficient by test-retest, split-half, or any usual reliability technique.[5] The tests were, however, administered to a convenience sample of 54 fifth- and sixth-grade pupils. These pupils were matched as nearly as possible on IQ and arithmetic achievement scores, resulting in 27 pairs of pupils. The scores of the matched pupils were then correlated for each test, with resulting correlations of .69 for the Written Heuristic Transfer test and .61

[5] For a discussion of problems associated with assessing reliability of instruments of this type, the reader is referred to Thorndike (1951, pp. 615–616).

for the Oral Heuristic Transfer test. These coefficients were used as crude estimates of the reliability of these instruments.

Pupil attitude toward arithmetic was assessed by two attitude scales administered in the pretest series and again in the posttest series to *Ss* in both treatments. The first of these (Semantic Differential Attitude Scale) was constructed along the principles of semantic differentiation developed by Osgood, Suci, and Tannenbaum (1958). All of the scales used in scoring this test were scales which received repeated high loadings on the evaluative factor in several of Osgood's studies (Osgood et al., 1958, ch. 2). The second scale (Statement Attitude Scale) was adapted from a similar instrument used by Umemoto and Haslerud (1955) and elicited expressions of pupil agreement or disagreement with both favorable and unfavorable statements concerning arithmetic. Since the two scales were judged to measure slightly different dimensions of attitude,[6] each *S*'s scores on the two scales were summed into a total attitude score (Total Attitude Scale).

In addition to these criterion measures, two instruments were used to assess the degree to which teachers adhered to the prescribed teaching model in each treatment. The first of these (Pupil Perception of Teaching Behavior) was a forced-choice questionnaire which elicited pupil responses to statements about teaching-behavior characteristics of their teacher. The statements reflected the five distinguishing characteristics of teaching behavior which were used in defining the expository and discovery teaching models described earlier. Each statement could be classified on the basis of which of the five "teaching-behavior characteristics" it represented. In addition, the statements could be classified into items which, if answered affirmatively, would typify pupil perception of "low discovery" teaching behavior, and items which, if answered affirmatively, would typify pupil perception of "high discovery" teaching behavior.

The scoring system, adapted from a similar pupil questionnaire reported by Shaw (1963, p. 3), was scaled so that a teacher-index score of 100 reflected pupil perception of maximum "high discovery-low expository" teaching behavior, while a teacher-index score of 0 reflected pupil perception of maximum "low discovery-high expository" teaching behavior.

In order to assess pupil perception of teaching behavior on the discovery-expository dimension, both before and after the experimental period, this instrument was administered in both the pre- and posttest series. The pretest scores of necessity reflected pupils' perceptions of the teacher's typical behavior in teaching arithmetic *prior to* the experimental period. It was predicted that typical teaching behavior, as measured by the index scores, would be found to combine some elements of both discovery and expository teaching but would resemble most closely the expository teaching model.

The posttest index scores obtained by use of this instrument should have reflected, in some measure at least, pupils' perceptions of the teacher's typical behavior in teaching arithmetic *during* the experimental period. Inasmuch as the rating scale was constructed so that the "pure discovery" model would be ranked as 100 on the scale and the "pure expository" model would be ranked as 0, it was predicted that the pre- to posttest gains in teacher index scores would show the following trends: (*a*) a positive pre- to posttest gain in Treatment D, and (*b*) a negative pre- to posttest gain in Treatment E.

The second instrument used to assess teacher fidelity to the prescribed models of teaching was a rating scale (Observer Rating of Teaching Behavior) used to rate teaching behavior through classroom observation and rating from audio-tape recordings. All raters were thoroughly trained in the selected mathematical concepts and in both treatment models during the training program described earlier. The percentage of rater agreement on this instrument was defined as the percentage of the total pairs of ratings assigned by two or more raters (rating the same teacher on the same lesson) which were in perfect agreement. Multiple rating of a lesson consisted of three modes: (*a*) instances when two observer-raters simultaneously visited the same classroom; (*b*) instances when an observer-rater rated a lesson which was also audiotaped, transcribed, and rated by a tape-rater; and (*c*) instances when a tape transcription was rated by all raters. The overall rater agreement obtained from these methods was .76. More definitive information on percentage of rater agreement for each mode is contained in the original report of the study (Worthen, 1965).

Ratings on the Observer Rating of Teaching Behavior scale were subdivided into two categories: (*a*) ratings on those items used to differentiate between discovery and expository teaching, and (*b*) ratings on items used to assess general teaching behavior. Data yielded by items in the latter category are not discussed in this report, but are included in the previous reports of this research. The items in the former category reflected the five teaching behavior characteristics used to define the treatment models. Comparisons of the mean ratings for each teacher in each treatment on these items were used to determine teacher fidelity to these prescribed models. Scoring of the rating scale was designed so that perfect adherence to the discovery model would have theoretically resulted in a mean teacher rating of 5.0, while perfect adherence to the expository model would theoretically have resulted in a mean teacher rating of 1.0.

The Pintner Intermediate Test; Form A (IQ) and the Metropolitan Achievement Test, Tests 5 and 6 (arithmetic computation and arithmetic problem-solving) were used as measures of group comparability.

[6] The intercorrelation (Pearson r) between them was .54 on a consecutive day test-retest.

TABLE 2
MEAN TEACHER RATINGS ON OBSERVER RATING OF TEACHING BEHAVIOR SCALE: BY TEACHER AND TREATMENT

Treatment	Teacher								Total treatment	Perfect score for theoretical model
	1	2	3	4	5	6	7	8		
D	4.76	4.68	4.73	4.32	4.38	4.53	4.73	4.45	4.56	5.00
E	1.40	1.32	1.20	1.56	1.30	1.56	1.20	1.68	1.41	1.00

RESULTS

Analysis of Teaching Behavior

Many investigations in which each teacher presented the learning task by two or more methods have been subject to the criticism that the teachers were unable to vary their teaching behavior sufficiently to effect a real test of the various treatment models. In order to obviate such criticism of the present study, two instruments described earlier (Observer Rating of Teaching Behavior and Pupil Perception of Teaching Behavior) were used to gather data on teaching behavior which might be characterized as "discovery" or "expository" in nature. These data were analyzed by use of the standard analysis of variance.

Summary of Observer Rating of Teaching Behavior. Table 2 shows the mean rating by teacher and treatment for this instrument.

Four analyses of variance were carried out on these data. The first analysis of variance compared mean teacher ratings in Treatment D with mean teacher ratings in Treatment E and yielded a highly significant difference between the two experimental treatments. This analysis confirmed the notion that the instruction given to both groups was consistently dissimilar.

The second analysis of variance compared the proximity of the obtained mean teacher ratings in each treatment to the perfect score for each of the theoretical teaching models. No significant differences were found in this comparison, thus substantiating the idea that instruction in both treatments followed the prescribed teaching models equally well.

The mean teacher ratings were also used to compare teacher differences within treatments. One analysis of variance compared the mean ratings for each teacher in Treatment D, while the second analysis compared the differences among the mean teacher ratings in Treatment E. No significant differences between teachers were found with either of these analyses.

Summary of Pupil Perception of Teaching Behavior. This instrument was used in an attempt to assess pupil perception of teaching behavior on the discovery-expository dimension both before and after the experimental period. The mean pretest index score for each teacher in each treatment is reported in Table 3. Although the prediction that typical teaching behavior would resemble most closely the expository teaching model was borne out (total treatment pretest means of 42.9 and 44.5), the pretest means approached rather closely the theoretical mean of 50 which would reflect discovery and expository teaching in equal proportions.

This questionnaire was scaled so that the pre- to posttest gain score for each teacher in each treatment could be used as an index of the teacher's adherence to the teaching model. In Treatment D, high fidelity to that model of teaching should have resulted in a positive pre- to posttest gain score. In Treatment E, high fidelity to that model of teaching should have resulted in a negative gain score. The mean pre- to posttest gains for each teacher in Treatments D and E are also presented in Table 3 and show a definite gain for each experimental treatment group in the predicted direction.

An analysis of variance which compared mean teacher gain scores in the two treatments revealed a highly significant difference ($F = 25.59$, $df = 1/398$, $p < .001$).

TABLE 3

MEAN PRETEST, POSTTEST, AND GAIN SCORES, AND STANDARD DEVIATIONS ON PUPIL PERCEPTION OF TEACHING BEHAVIOR: BY TEACHER AND TREATMENT

Treatment	Pretest, posttest, or gain	Teacher 1	2	3	4	5	6	7	8	Total treatment
D	Pre									
	M	35.4	38.6	43.2	47.9	43.8	39.7	46.9	47.7	42.9
	SD	19.3	17.8	18.6	22.9	21.7	18.4	19.0	23.8	20.8
	Post									
	M	43.5	45.6	56.0	54.5	48.5	46.8	50.1	57.4	50.3
	SD	22.1	21.0	17.5	21.7	24.0	17.4	17.1	22.9	21.2
	Gain									
	M	8.1	7.0	12.8	6.6	4.7	7.1	3.2	9.6	7.4
	SD	13.8	8.8	12.0	12.3	13.6	12.6	8.6	12.6	12.2
E	Pre									
	M	44.0	42.3	42.8	45.7	45.0	40.8	51.8	43.5	44.5
	SD	26.2	19.7	20.4	18.7	23.6	16.8	16.6	19.9	20.5
	Post									
	M	42.6	37.5	44.4	44.0	44.6	36.0	47.3	46.9	42.9
	SD	21.3	17.2	16.1	19.9	21.1	17.2	16.8	22.0	19.6
	Gain									
	M	−1.4	−4.8	1.6	−1.7	−.4	−4.8	−4.5	3.5	−1.6
	SD	12.8	8.6	8.4	8.6	12.6	11.4	9.8	9.3	10.6

These data were interpreted as further evidence that the teachers varied their behavior sufficiently to effect a real test of the two teaching models. No significant differences were found among teacher pre- to posttest gain scores within either treatment.

Summary of Tests of Hypotheses

Because of the noncomparability of the treatment groups on some pretreatment measures, statistical controls were imposed in all intertreatment data analyses (except analyses of teaching-behavior data discussed previously) by use of a two-way teacher-by-treatment analysis of covariance.

The choice of covariates was determined by an examination of the intercorrelations on all measures and variables. On this basis, IQ, arithmetic computation, and arithmetic problem-solving were used as constant covariates in the analysis of each dependent variable. Pretest scores were used as additional covariates in analysis of the posttest of each instrument administered in both the pre- and posttest series. Posttest scores on the Concept Knowledge Test were used as an additional covariate in the analysis of the Concept Retention and Concept Transfer tests.

Significant F ratios for between-teacher effects and Teacher × Treatment interaction were yielded by the analysis of each criterion measure. No completely satisfactory explanation could be given for these results, although three plausible explanations are offered below.

1. Variables of teaching behavior and personality too fine to be detected by the gross measures of teaching behavior used in this study were operative for all teachers.

2. Certain teacher personalities were more compatible with one of the experimental treatments than with the other, thus resulting in Teacher × Treatment interaction effects.

3. The variation between schools in the number of units completed could also be a highly plausible explanation of the significant between-teacher F ratio yielded by all measures used in this study.

No further attempt to explain these findings is given here. Only the results yielded by comparisons between Treatments D and E are presented. Posttest means and standard deviations for each criterion measure are shown in Table 4 along with the post-

TABLE 4

POSTTEST MEANS AND STANDARD DEVIATIONS; POSTTEST MEANS ADJUSTED BY COVARIANCE AND STANDARD ERRORS OF ADJUSTED POSTTEST MEANS

Measure	Treatment	Posttest		Adjusted Posttest	
		M	*SD*	*M*	*SE*
Concept Knowledge Test	D	52.5	17.1	51.5	.78
	E	54.0	18.0	54.2	.78
Concept Retention Test 1	D	48.9	20.9	42.7	.86
	E	46.7	21.3	40.1	.91
Concept Retention Test 2	D	48.9	20.7	a	a
	E	46.1	23.2		
Concept Transfer Test	D	34.0	19.5	30.3	.82
	E	31.5	18.3	28.2	.87
Negative Concept Transfer Test	D	10.8	6.2	10.6	.38
	E	10.6	6.0	10.7	.38
Semantic Differential Attitude Scale	D	16.3	7.9	16.7	.46
	E	16.8	8.4	16.4	.46
Statement Attitude Scale	D	30.0	21.6	32.7	.99
	E	32.2	20.1	31.1	.98
Total Attitude Scale	D	46.3	27.9	49.6	1.28
	E	49.1	26.7	47.3	1.26
Written Heuristic Transfer	D	21.1	11.4	a	a
	E	18.8	10.7		
Oral Heuristic Transfer	D	18.5	11.5	a	a
	E	15.9	11.6		

[a] The computer program used in the analyses of covariance of Concept Retention Test 2 and both heuristic transfer tests did not give directly adjusted means.

test means adjusted by covariance and standard errors of the adjusted means.

Initial learning. The data yielded by the Concept Knowledge Test did not support the hypothesis that Treatment D would produce superior results on an initial learning test. On the contrary, these data showed Treatment E to produce significantly better results ($p < .01$) than Treatment D on the initial learning criterion test.

Retention. The hypothesis that Treatment D would produce superior results to Treatment E on a retention test given 5 weeks and 11 weeks after instruction was supported by the evidence yielded by an analysis of the Concept Retention Test scores ($p < .05$ for the first administration and $p < .025$ for the second).[7]

Concept transfer. The data yielded by the Concept Transfer Test lent slight support ($p < .08$) to the hypothesis that pupils in Treatment D would show greater ability to transfer the concepts learned during instruction than would pupils in Treatment E.

Negative concept transfer. There was no support in the data yielded by the Negative Concept Transfer Test for the hypothesis that Treatment D would produce less negative transfer than Treatment E. Rather, it was found that there were no differences in negative transfer between the treatments.

Attitude. Of the three possible comparisons between Treatments D and E on measures of attitude, none reached significance at a minimum acceptable level. The hypothesis that Treatment D would produce superior results to Treatment E on attitude measures was not supported by the data.

Transfer of heuristics. The hypothesis that Treatment D would produce superior results to Treatment E on tests of pupil ability to transfer heuristics was supported by the evidence yielded by analyses of scores from both the Written Heuristic Transfer Test ($p < .05$) and the Oral Heuristic Transfer Test ($p < .025$).

Table 5 summarizes the analyses of covariance which yielded the above results.

DISCUSSION AND CONCLUSIONS

Teaching Behavior

Of most importance for the interpretation of the results of this study was the clear-cut evidence that *Ss* in the two

[7] The Concept Knowledge Test represents the summation of four discrete subtests, each of which was administered immediately upon completion of the corresponding subsection of instructional materials. This resulted in a series of four staggered posttests given approximately 8, 6, 4, and 3 weeks prior to the first administration of the Concept Retention Test. The four subscores were summed to yield a total Concept Knowledge Test score. The average delay between administration of these subtests and the first administration of the Concept Retention Test was slightly over 5 weeks. The second administration of the Concept Retention Test came 6 weeks after the first. Therefore, the average time between the subtests and the second retention test was slightly over 11 weeks. The reader may note that the second administration of the Concept Retention Test was not reported in previous reports of this study (Della-Piana, Eldredge, & Worthen, 1965; Worthen, 1965). Both reports were written to meet deadlines which came before the second administration of the Concept Retention Test. Data from this second retention test were analyzed after the earlier reports had been printed.

treatments received instruction by two consistently different methods of teaching, each of which closely paralleled the particular model prescribed. Although the necessity of experimental controls may have precluded either method from reaching its optimum power, this factor, if present, was equally operative in both treatments.

Test of Hypotheses

In general, the findings of this study seem to support many of the claims made by proponents of discovery methods. The most dramatic finding was the rather startling reversal in rank of Treatments D and E between the administration of the Concept Knowledge posttest and the first administration of the Concept Retention Test 5 weeks later. Although Treatment E was significantly superior to Treatment D on the tests of initial learning ($p < .01$), the retention test given after an average 5-week delay showed Treatment E not only to have lost this initial superiority but also to have been surpassed by Treatment D. The *S*s taught by the discovery method were able to retain significantly more material ($p < .05$) over the intervening period, notwithstanding the fact that they had evidenced knowledge of significantly less material than the Treatment E group on the test of initial learning. The second administration of the Concept Retention Test 6 weeks after its first administration showed Treatment D to have maintained this superiority over Treatment E ($p < .025$). This finding strongly suggests that presentation of mathematical concepts to sixth-grade pupils through discovery sequencing causes the learner to integrate the content conceptually in such a manner that he can retain it more readily than if the concepts had been presented to him in an expository sequence.

Another finding which clearly favors Treatment D is that dealing with *S*s' acquisition of a problem-solving set. In light of the evidence yielded by both the Written Heuristic Transfer and the Oral Heuristic Transfer tests, it seems reasonable to conclude that learning by discovery techniques significantly increases pupil ability to use discovery problem-solving approaches in new situations, both those which require paper and pencil application and those which involve verbal presentation by the teacher. Treatment D was shown to be significantly superior to Treatment E on both of these dimensions in the present study.

TABLE 5
SUMMARY OF ANALYSES OF COVARIANCE OF CRITERION MEASURE POSTTEST SCORES BETWEEN TREATMENTS D AND E

Measure	df_1	df_2	F	Direction
Concept Knowledge Test	1	412	7.435****	D < E
Concept Retention Test 1	1	412	3.918**	D > E
Concept Retention Test 2	1	412	5.868***	D > E
Concept Transfer Test	1	412	3.089*	D > E
Negative Concept Transfer Test	1	413	.098	
Semantic Differential Attitude Scale	1	412	.161	
Statement Attitude Scale	1	412	1.173	
Total Attitude Scale	1	412	2.057	
Written Heuristic Transfer	1	413	5.004**	D > E
Oral Heuristic Transfer	1	413	5.720***	D > E

* $p < .08$.
** $p < .05$.
*** $p < .025$.
**** $p < .01$.

Treatment D also seems superior to Treatment E in terms of transfer, although this finding is somewhat tenuous. It was the experimenter's opinion that the Concept Transfer Test was much too difficult for *S*s involved and that this factor reduced the possibility of finding more significant differences between the treatments. Random errors of measurement, due to difficulty and resultant guessing of answers, could act to increase the analysis of covariance error term, thus making it more difficult to obtain a significant F ratio. The obtained F ratio favored Treatment D over Treatment E at a marginal level of significance ($p < .08$), and the experimenter would speculate that modifications of the instrument to reduce the random error of measurement would result in more highly significant differences in favor of Treatment D.

The question of relative practicality of discovery and expository teaching in terms of time consumption has been raised by Ausubel (1961, 1964). It should be noted that the controls established in this study can be used to answer this question with reference to fifth- and sixth-grade pupils.

Each teacher equalized the length and number of daily learning periods for each of his two classes. This resulted in some variance among schools (due to class scheduling), but no variance between treatments when summed across schools. The results indicate that the discovery method need not be more time consuming than the expository method of instruction at this age level. When given an equal amount of time to work on the learning task, *S*s in Treatment D proved superior to *S*s taught by Treatment E in the majority of intertreatment comparisons.

Implications for future research. It was noted earlier that programmatic research dealing with various discovery-expository variables of task presentation should be initiated. In addition to a continuation of research in which sequence characteristics of the learning task are manipulated, the present research design and instructional materials might be modified to explore the relative effects of various types and amounts of guidance along the discovery-expository dimension. Studies could be designed in which the present learning task could be modified to compare discovery methods in which the verbal factor was varied from verbal to nonverbal discovery. Interrelationships among these relevant variables might then be explored.

Implications for educational practice. Any generalizations based on findings of this study must take into account the particular teachers, experimental sample, instructional procedures, instructional materials, and criterion measures used. In addition, without the programmatic research suggested above, any conclusions drawn on the basis of this single study must be tentative at best.

Conversely, this study was conducted under carefully controlled conditions which were judged to approximate normal classroom conditions with respect to all dimensions except those specifically varied for experimental purposes. Because of the relatively large time sample, the nature of the learning task, and the large number of *S*s used, it would seem that the results can be generalized, at least to innovative teaching with similar *S*s and subject matter content, with a relatively high degree of confidence. Within this context, it is the experimenter's opinion that, pending further programmatic research, this study holds the following implications for educational practice:

1. To the extent that pupil ability to retain mathematical concepts and pupil ability to transfer heuristics of problem solving are valued outcomes of education, discovery sequencing should be an integral part of the methodology used in presenting mathematics in the elementary classroom.

2. To the extent that immediate recall is a valued outcome of education, expository sequencing should be continued as the typical instructional practice used in the elementary classroom.

REFERENCES

Ausubel, D. P. Learning by discovery: Rationale and mystique. *Bulletin of National Association of Secondary School Principals,* 1961, **45,** 18–58.

Ausubel, D. P. Some psychological and educational limitations of learning by discovery. *Arithmetic Teacher,* 1964, **11,** 290–302.

Beberman, M. *An emerging program of secondary school mathematics.* Cambridge: Harvard University Press, 1958.

Bellack, A. A., Davitz, J. R., Kleibard, H. M., & Hyman, R. T. *The language of the classroom: Meanings communicated in high school teaching.* New York: Bureau of Publications, Teachers College, Columbia University, 1963.

Bruner, J. S. The act of discovery. *Harvard Educational Review,* **1961,** 21–32.

Corman, B. R. The effect of varying amounts and kinds of information as guidance in problem solving. *Psychological Monographs,* 1957, **71**(2, Whole No. 431).

Craig, R. C. *The transfer value of guided learning.* New York: Bureau of Publications, Teachers College, Columbia University, 1953.

Craig, R. C. Directed vs. independent discovery of established relationships. *Journal of Educational Psychology,* 1956, **47,** 223–224.

Della-Piana, G. M. Searching orientation and concept learning. *Journal of Educational Psychology,* 1957, **48,** 245–253.

Della-Piana, G. M., Eldredge, G., & Worthen, B. R. *Sequence characteristics of text materials and transfer of learning. Part 1. Experiments in discovery learning.* Salt Lake City: Bureau of Educational Research, University of Utah, 1965.

Gagné, R. M., & Brown, L. T. Some factors in the programming of conceptual learning. *Journal Experimental Psychology,* 1961, **62,** 313–321.

Hendrix, G. A new clue to transfer of training. *Elementary School Journal,* 1947, **48,** 197–208.

Kersh, B. Y. The adequacy of 'meaning' as an explanation for the superiority of learning by in-

dependent discovery. *Journal of Educational Psychology,* 1958, **49**, 282–292.

KERSH, B. Y. The motivating effect of learning by directed discovery. *Journal of Educational Psychology,* 1962, **53**, 65–71.

KITTELL, J. E. An experimental study of the effect of external direction during learning on transfer and retention of principles. *Journal of Educational Psychology,* 1957, **48**, 391–405.

McDONALD, F. J. Meaningful learning and retention: Task and method variables. *Review of Educational Research,* 1964, **34**, 530–544.

OSGOOD, C. E., SUCI, G. J., & TANNENBAUM, P. H. *The measurement of meaning.* Urbana: University of Illinois Press, 1958.

RAY, W. S. Pupil discovery vs. direct instruction. *Journal of Experimental Education,* 1961, **29**, 271–280.

SHAW, J. A. A students' inventory of teaching-learning processes: A manual for forms C. & D. Colorado State College, 1963. (mimeo)

SUCHMAN, J. R. *The elementary school training program in scientific inquiry.* University of Illinois, 1962.

SWENSON, E. J. Organization and generalization as factors in learning, transfer, and retroactive inhibition. In University of Minnesota Studies in Education, No. 2, *Learning theory in school situations.* Minneapolis: University of Minnesota Press, 1949. Pp. 9–39.

THORNDIKE, R. L. Reliability. In E. F. Lindquist (Ed.), *Educational measurement.* Menasha, Wisconsin: Banta, 1951. Pp. 560–620.

UMEMOTO, T., & HASLERUD, G. Responses and attitudes of American and Japanese children to number situations: A comparative study. Memoirs of the Kyoto University College of Education, 1959, **5**, 69–96. (Trans. by M. M. Tatsuoka. Copies of the translation available from University of Illinois Committee on School Mathematics Project.)

WITTROCK, M. C. The learning by discovery hypothesis. In L. S. Shulman & E. R. Keisler (Ed.), *Learning by discovery: A critical appraisal.* Chicago: Rand McNally, 1966. Pp. 33–75.

WORTHEN, B. R. Discovery vs. expository classroom instruction: An investigation of teaching mathematics in the elementary school. Unpublished masters thesis, University of Utah, 1965.

(Received January 3, 1967)

JOURNAL OF EDUCATIONAL PSYCHOLOGY

Volume 59 | April 1968 | Number 2

CONTENTS

© 1968 by the American Psychological Association, Inc.

Journal of Educational Psychology
1968, Vol. 59, No. 2, 59-68

LIBRARY

INDUCTION AND TRANSFER OF SEARCH SETS

BRUCE W. TUCKMAN
Rutgers—The State University

JAMES HENKELMAN, GERALD P. O'SHAUGHNESSY,
AND MILDRED B. COLE[1]
University of Maryland

This study investigated the effects of appropriate and inappropriate practice experiences on students' tendency to search for and find shortcut solutions to problems. The strategy of looking for, and skill in finding, a shortcut were termed a search set. In the 1st of 3 experiments, with criterion problems highly similar to the practice problems, Ss having search experiences were more inclined to search for and find shortcut solutions than those having nonsearch experiences. In the 2nd and 3rd experiments, with the criterion problems more dissimilar from the practice problems, Ss having search experiences were more inclined to search for a shortcut than those having nonsearch experiences, but were notably unsuccessful in finding the shortcut solution. Thus, the strategy of search could be made to transfer more readily than the skill required to search successfully.

The purpose of this study was to explore the inductive approach to set learning in terms of the conditions under which it takes place and the conditions under which it transfers. In particular the study dealt with the search set which included both the *strategy* or inclination to search for an "elegant" or shortcut solution to an educationally relevant problem and the *skill* or knowledge to carry out the search successfully and discover the appropriate solution.

Learning-set formation as identified by Harlow (1949, 1959) refers to the phenomenon of transfer of training between many problems of a single class. The organisms, in Harlow's case rhesus monkeys, were trained so that they could solve problems that were common in "principle" but different in specific stimulus content than the problems used during training. Such learning required that certain error factors such as stimulus preference or position preference be overcome during training; the monkeys had to learn to *discriminate* between cues which led to an incorrect solution and a class of cues that led to a correct solution.

Gagné and Paradise (1961) adopted the term "learning set" to refer to specific "knowledges," that is, principles or concepts, subordinate to some final learning task. In the Gagné and Paradise sense, the term learning set referred to learning that had taken place and could then be applied to further, higher-order, learning, and did not refer to the "being set" or "being oriented" toward a particular problem-solving approach as implied in the Harlow sense of the term.

Another track pursued in the study of problem solving and concept learning has been the strategy approach as used by Goodnow (1955), Goodnow and Pettigrew (1955), and Bruner, Goodnow, and Austin (1956). Goodnow and Pettigrew (1955) defined a strategy as a consistent way of deciding or responding. These authors rejected the notion of describing their particular problem-solving paradigm using the concept of strategy by itself. They chose to employ the term *learning set* to

[1] The authors would like to express their appreciation to the members of the University of Maryland Mathematics Project who provided support for the study, namely: John R. Mayor, Henry H. Walbesser, Sherman B. Vanaman, John R. Kolb, Harry G. Crowcroft, William L. Gray, and Ingrid Weise. Without their help and the support of the Mathematics Project, this study would not have been possible. Portions of this study were read at the meetings of the American Educational Research Association in New York, February 1967.

apply to the knowledge or learning that had taken place (in the Gagné & Paradise, 1961, sense), and the term *strategy* was restricted to consistent use of, or sensitization to, search in a particular fashion.

The term *search set* is offered to include both learning set and strategy concepts in a problem-solving situation. Individuals can be set to approach a problem in a particular way but will only be successful in this approach if they have acquired the appropriate knowledge. Having a search set means being set to respond to a meaningful stimulus configuration in a prescribed manner and having sufficient knowledge to successfully follow this prescription. It is not inconceivable that the skill and strategy components may not be present in equal amounts.

PROBLEM

The purpose of the present research was to attempt to induce a search set in individuals through experimentally controlled prior experience and to determine the conditions which would allow the search set to transfer. Ideally, a set has unlimited transfer possibilities (Gagné, Mayor, Garstens, & Paradise, 1962), and this experiment tested the accuracy of such a statement.

Goodnow and Pettigrew (1955) demonstrated that the learning set-strategy package could be induced by controlled prior experiences. Individuals who were exposed to a 100:0 pattern learned a later 100:0 pattern or 0:100 pattern more quickly following a 50:50 exposure than subjects (*S*s) who had never experienced the initial 100:0 pattern.

Luchins (1942), in a classic series of studies, investigated the effect of *Einstellung*, a reasonable synonym for set. He found that prior exposures to problems having a common solution route produced extreme rigidity in the continued use of that solution route even when other, more direct, routes were available. The undesirability of *Einstellung* is not clearly demonstrated in Luchin's work although that appeared to be his intention. In some cases the *Einstellung* solution was a correct one albeit a longer one to accomplish. However, if the alternative to the *Einstellung* solution was no solution, then the *Einstellung* solution takes on greater value than that attributed to it by Luchins. In fact, Luchins showed that increasing the time pressure increased the likelihood of an *Einstellung* solution, implying that *S*s valued a solution over no solution.

In this research, the emphasis is the reverse of Luchins', that is, to induce the direct or "elegant" solution route through controlled prior experience in an effort to overcome a more obvious or "inelegant" approach which was available and had been well-practiced by *S*. That is, rather than attempting to induce rigidity of the inelegant or unnecessarily long solution route as Luchins did, the purpose here was primarily to induce and teach a search set that represented a break with a more traditional solution route in the direction of a more elegant solution. The Luchins' approach of inducing rigidity through inappropriate experience was included as a control condition against which the effects of appropriate experience could be evaluated. In keeping with Luchins' attempt to alter behavior by instruction, two instruction conditions, one appropriate and one inappropriate, were also included in this experiment. Finally, an attempt was made to see whether experiences appropriate for a particular search set would allow for that set to be transferred to situations for which its appropriateness was not so obvious.

The mathematician's utilization of the terms "elegant" and "inelegant" solutions to discriminate between alternative correct approaches on the basis of their cleverness and speed are appropriate descriptions for the search set and its antithesis as used in this research. Befitting the use of these terms, a mathematical problem was the vehicle for studying search-set induction.

Finally, the relevance of the problem to educational settings lies in the fact that the educator is anxious to foster learning by discovery (i.e., induction) in the classroom

since, as the findings of Gagné and Brown (1961) and Katona (1940) show, such learning provides for greater transfer of the concepts and principles to new but related problems. Transfer and transferability are clearly goals of the educational process.

Hypotheses

The following hypotheses were offered:

1. Individuals having search experiences will be more inclined to seek and discover a search solution on a problem highly similar to that experienced than those having nonsearch experiences.
2. Individuals given a search instruction will be more inclined to seek and discover a search solution than those not so instructed.
3. Individuals having search experiences will be more inclined to seek and discover a search solution on a problem only moderately similar in principle to that experienced as compared to individuals having nonsearch experiences.

It is anticipated that experiences in solving problems which have been structured so that a particular class of search solutions works will lead to learning by discovery of that class of solutions (this has been demonstrated by Gagné & Brown, 1961) as well as disposing persons toward the use of a search strategy (cf. Goodnow & Pettigrew, 1955). Such learning and strategy adoption as a result of solving search problems should show a range of transferability to be of practical significance in an educational context. While the use of an instruction or suggestion will not provide for set learning, it may be equally strategy-inducing as structured experiences. (Studies such as that of Pepitone & Reichling, 1955, have successfully induced a cohesiveness set via instruction.) Induction by instruction should yield more successful search than no induction or nonsearch induction by instruction.

EXPERIMENT I

Subjects

Participants in Experiment I were 262 female undergraduate students at the University of Maryland. Participation occurred as part of regular undergraduate mathematics classes.

The Task

The task was a 4 × 6 matrix of two-digit numbers as follows that were to be added:

18	18	18	18	18	18
10	10	10	10	10	10
24	24	24	24	24	24
79	79	79	79	79	79

Those problems identified as *search problems* could be solved by methods other than adding all 24 numbers. The example given above represents a search problem that could be solved (i.e., the sum could be found) by adding the numbers in the first column and multiplying by 6 since all the columns are identical. Other search problems featured two columns of numbers repeated three times each; others featured a row repeated four times although the order of the numbers was different from row to row. The common feature of search problems was that a shortcut method could be used to obtain a sum rather than adding all 24 numbers.

A *nonsearch problem* was a matrix of 24 numbers that could only be summed by adding all 24 numbers; no pattern existed so that no shortcut solution could be used. Search and nonsearch problems were equated in pairs in terms of the amount of time required to solve each by summing all 24 numbers. However, solving search problems via the shortcut required considerably less time once the shortcut was discovered.

Procedure

The *S*s were randomly assigned to one of the eight conditions shown in Figure 1. The eight conditions run are described below:

Cell 1. *S*s were given a booklet containing four problems, each of which could be solved by a shortcut. The first three constituted search practice; the fourth was a search criterion problem.

Cell 2. *S*s were given a booklet containing four problems, none of which could be solved by a shortcut. The first three constituted nonsearch practice; the fourth was a nonsearch criterion problem.

Cell 3. *S*s were given a booklet containing four problems. The first three could not be solved by a shortcut (same problems as Cell 2) and constituted nonsearch practice. The fourth problem could be solved by a shortcut and constituted a search criterion problem (same criterion problem as Cell 1).

Cell 4. *S*s were given a booklet containing four problems. The first three could be solved by a shortcut (same problems as Cell 1) and constituted search practice. The fourth problem could not be solved by a shortcut and constituted a nonsearch criterion problem (same criterion problem as Cell 2).

(Thus, Cells 1 and 2 represent prior experience appropriate for the criterion problem while Cells

	Appropriate prior experience	Inappropriate prior experience	Appropriate suggestion	Inappropriate suggestion	No intervention
Search criterion problem	Cell 1 3 search problems— search criterion problem $N = 35$ $M = 60$ SD = 6.0	Cell 3 3 nonsearch problems— search criterion problem $N = 35$ $M = 85$ SD = 6.0	Cell 5 search suggested on search criterion problem $N = 32$ $M = 108$ SD = 7.0		Cell 7 search criterion problem $N = 32$ $M = 111$ SD = 9.5
Nonsearch criterion problem	Cell 2 3 nonsearch problems— nonsearch criterion problem $N = 33$ $M = 94$ SD = 4.9	Cell 4 3 search problems— nonsearch criterion problem $N = 34$ $M = 115$ SD = 5.5		Cell 6 search suggested on nonsearch criterion problem $N = 31$ $M = 135$ SD = 6.6	Cell 8 nonsearch criterion problem $N = 30$ $M = 134$ SD = 6.3

FIG. 1. Design of Experiment I and mean times to completion in seconds (and standard deviations) for each cell.

3 and 4 represent prior experience inappropriate for the criterion problem.)

Cell 5. *S*s were given a single problem which could be solved by a shortcut constituting a search criterion problem. Appearing above the problem was the statement, "Problems such as this can often be solved using a shortcut." This represents appropriate suggestion.

Cell 6. *S*s were given a single problem which could not be solved by a shortcut constituting a nonsearch criterion problem. At the top of the page appeared the same statement as in Cell 5. This represents inappropriate suggestion.

Cell 7. *S*s were given a single search criterion problem.

Cell 8. *S*s were given a single nonsearch criterion problem.

The search criterion problem used in Cells 1, 3, 5, and 7 and the nonsearch criterion problem used in Cells 2, 4, 6, and 8 appear below.

Search criterion problem						Nonsearch criterion problem					
13	28	49	49	64	13	61	59	68	89	32	99
28	49	64	13	49	64	27	72	68	21	67	63
49	13	13	64	28	28	78	95	49	58	33	16
64	64	28	28	13	49	71	9	9	40	34	38

Identifying Search Behavior

All *S*s timed themselves on all problems and recorded their time to completion in seconds on the page of the problem. All scratch work was also done on this page. (No feedback was given in any instance.) Since search problems could be solved in considerably less time than nonsearch problems once the shortcut was discovered, the effects of prior experience or suggestion was assessed in terms of the time to problem completion (i.e., time taken to obtain a sum of the numbers). Specifically, Hypothesis 1 was tested by comparing mean time to solution on the criterion problem for those groups having appropriate prior experience (Cells 1 and 2) to those groups having inappropriate prior experience (Cells 3 and 4). It was reasoned that:

1. Cell 1 *S*s, if affected by the experience, would be set to search for a shortcut solution, have skill in finding it, and subsequently complete the criterion problem in a minimum amount of time;

2. Cell 2 *S*s, if affected by the experience, would be set *not* to search for a shortcut solution, and sum all the numbers in an intermediate amount of time (the shortcut route takes less time than straight summing);

3. Cell 3 *S*s, if affected by the experience, would be set *not* to search for a shortcut solution, and sum all the numbers in an intermediate amount of time;

4. Cell 4 *S*s, if affected by the experience, would be set to search for a shortcut solution, and not finding one since none existed, would end up solv-

ing the problem by summing—thus using more time than any of the preceding three cells.

Other comparisons were also made to assess the affects of practice and of appropriate and inappropriate suggestion.

RESULTS

Means and standard deviations for each of the eight cells of the experiment appear in Figure 1. The results of an analysis of variance of solution times for *S*s in the first four cells of the experiment indicates that *S*s having appropriate prior experience (Cells 1 and 2) took significantly less time to complete the criterion problem than *S*s having inappropriate prior experience (Cells 3 and 4) ($F = 15.8$, $df = 1/133$, $p <$ .01). A significant main effect of criterion problem was also obtained with the search criterion problem requiring less time to completion than the nonsearch criterion problem ($F = 30.6$, $df = 1/133$, $p < .001$).

Cell means were further compared using the Duncan multiple-range test (Winer, 1962) and all mean differences *except* that between Cells 2 and 3 were significant at the .01 level.

The fact that *S*s in Cell 1 took less time to completion than *S*s in Cell 3 indicates that the former were both more set and more skilled in searching as a result of their prior search experience. The fact that *S*s in Cell 2 took less time to completion than *S*s in Cell 4 indicates that *S*s in Cell 4 spent some time searching inappropriately as a result of their prior search experiences. These findings confirm the first hypothesis.

The data also show that the effects of suggestion, either appropriate or inappropriate, were effectively nil. Comparing Cell 5 (appropriate suggestion) with Cell 7 (no intervention) using a t-test, time to completion on the search criterion problem was no different for the two groups, indicating that the suggestion did not significantly affect the search set. Comparing Cell 6 (inappropriate suggestion) to Cell 8 (no intervention) showed that on a nonsearch criterion problem the difference in time to completion between the two groups was not significant. Thus, inappropriate suggestion did not serve to affect *S*s adversely. This finding negated any necessity for completing the two missing cells in the design which would have helped to assess the effects of suggestion, had there been any.

Two interesting and unexpected findings emerged. The comparisons of Cells 1, 2, 3, and 4 (all those that had any prior experience, whether appropriate or inappropriate) versus Cells 5, 6, 7, and 8 (those not having any prior experience) show the former to have taken significantly less time to complete the criterion problem ($t = 6.03$, $df = 264$, $p < .001$). This finding suggests that there is a general familiarity or practice effect that enhances the performance of *S*s on this type of a problem, independent of the set which is established by prior experience.

A second finding is also worthy of note. If the effects of inappropriate prior experience had been complete, the time to completion taken by *S*s in Cell 3 should have been identical to that taken for *S*s in Cell 2. There should not have been any differential reaction to the criterion problem itself since both cells had prior experience with nonsearch problems. However, as the data showed, *S*s in Cell 3 took less time to complete the problem than *S*s in Cell 2. While this difference did not approach significance it is, however, worthy of note. The significant main effect of criterion problem leads one to conclude that the shortcut in the search criterion problem is sufficiently visible to be found by some *S*s having nonsearch experience and would account for the Cell 3 effect. Furthermore, a comparison of Cells 7 and 8, neither of which had any experience, indicated that the search criterion problem as a stimulus array prompted search since Cell 7 *S*s took significantly less time to solution than Cell 8 *S*s ($t = 1.80$, $df = 60$, $p < .05$).

To determine with more exactness the kinds of behavior going on among *S*s in Cell 3, an attempt was made to analyze the protocols of *S*s in this cell and compare them to the protocols of *S*s in Cell 1. Specifically, an attempt was made to determine which *S*s actually carried out search or shortcut solutions as opposed to

TABLE 1

NUMBER OF SUBJECTS IN CELLS 1 AND 3 WHO ATTAINED SEARCH, EMERGENT SEARCH, AND NONSEARCH SOLUTIONS TO THE SEARCH CRITERION PROBLEM (EXPERIMENT I)

Solution	Cell 1	Cell 3	Total
Search	24	9	33
Emergent	7	15	22
Nonsearch	4	11	15
Total	35	35	70

those who did not carry out such solutions. A third group was also identified in which *S*s appeared not to begin with search solutions but to end up with search solutions, such solutions having emerged from their work on the search criterion problem. This analysis of protocols could not be done on Cells 2 and 4 since these two cells used a nonsearch criterion problem for which no search solution was possible. The purpose of the comparison between the cells was to determine the extent to which inappropriate prior experiences for Cell 3 *S*s had a complete versus a partial effect in terms of causing them to seek nonsearch solutions to the search criterion problem. This breakdown by type of solution utilized appears in Table 1. As the table shows, *S*s in Cell 1 used significantly more search solutions than *S*s in Cell 3 ($\chi^2 = 10.56$, $df = 2$, $p < .01$), thus supporting the time data. However, it can be seen that the effects of inappropriate prior experience on search criterion problem solution was incomplete since some of the Cell 3 *S*s immediately used the search solution while others discovered a search solution in an emergent fashion. From this one must conclude again that the search criterion problem itself, as a stimulus configuration, influenced the behavior of *S*s above and beyond the effects of prior experience provided in the experiment.

EXPERIMENT II

Procedure

Experiment II was identical in procedure to the first four cells of Experiment I except that different criterion problems were used in the second ex-

	Appropriate prior experience	Inappropriate prior experience
Search transfer criterion problem	Cell 1a 3 search problems—search transfer criterion problem $N = 50$ $M = 107$ SD = 6.9	Cell 3a 3 nonsearch problems—search transfer criterion problem $N = 52$ $M = 97$ SD = 5.2
Nonsearch criterion problem	Cell 2a 3 nonsearch problems—nonsearch criterion problem $N = 51$ $M = 99$ SD = 5.4	Cell 4a 3 search problems—nonsearch criterion problem $N = 47$ $M = 122$ SD = 4.9

FIG. 2. Design of Experiment II and mean times to completion in seconds (and standard deviations) for each cell.

periment. Cells 5, 6, 7, and 8 of the first experiment were dropped due to the limited results and notable practice effect obtained in the first experiment. The criterion problems were altered to test the third hypothesis, dealing with the transferability of search sets. Specifically, the new search criterion problem required the use of a shortcut rule which had not been encountered in the three prior experience problems and thus required transfer. A new nonsearch criterion problem was developed and equated in time to solution by adding with the search criterion problem. Both criterion problems appear below.

Search criterion problem						Nonsearch criterion problem					
13	28	49	49	64	13	61	59	68	89	32	99
28	49	64	13	49	64	27	72	68	21	67	63
13	28	64	28	49	13	78	95	49	58	33	16
49	64	28	64	13	28	71	9	9	40	34	38

The search criterion problem could be solved by using the following rule: Multiply each of the four numbers that appears (13, 28, 49, and 64) by 6, which represents the number of times each appears, and then sum the four products.

Participants in the second experiment were 200 undergraduate girls using the same procedure as was employed in the first experiment. These *S*s came from the same total population as those in Experiment I. As in Experiment I, *S*s were randomly assigned to conditions.

The four conditions employed in Experiment II appear in Figure 2. Briefly, these are: (*a*) prior search experience—search transfer criterion problem; (*b*) prior nonsearch experience—nonsearch criterion problem; (*c*) prior nonsearch experience—search transfer criterion problem; (*d*) prior search experience—nonsearch criterion problem.

Again, time to solution and analysis of solutions were used as the criterion measures.

RESULTS

Means and standard deviations for each of the four cells of Experiment II appear in Figure 2. The results of an analysis of variance of solution times shows that the effects of neither appropriate versus inappropriate experience nor search versus nonsearch criterion problem was significant but that the interaction of these two variables was significant ($F = 12.2$, $df = 1/196$, $p < .01$). Inappropriate practice appeared to facilitate performance on the search criterion problem while it adversely affected performance on the nonsearch criterion problem. It would seem that practicing the nonsearch approach was more effective a prerequisite to completing the search transfer criterion problem than was practicing the search approach. This finding is opposite to that in the first experiment and leads to the rejection of the second hypothesis.

TABLE 2

NUMBER OF SUBJECTS IN CELLS 1a AND 3a WHO ATTAINED SEARCH AND NONSEARCH SOLUTIONS TO THE SEARCH CRITERION PROBLEM AND MEAN TIME TO SOLUTION FOR EACH SUBGROUP (EXPERIMENT II)

Solution	Cell 1a		Cell 3a		Total subjects
	Subjects	*M*	Subjects	*M*	
Search	14	79	3		17
Nonsearch	36	118	49	97	85
Total	50		52		102

In order to shed more light on this turn of events, protocols were examined and frequencies of search solutions and nonsearch solutions for *S*s in Cells 1a and 3a were identified.[2] These data appear in Table 2. Apparently, relatively few *S*s in either cell attained a search solution to solve the problem. However, the number of *S*s in Cell 1a attaining a search solution significantly exceeded that of Cell 3a ($\chi^2 = 7.64$, $df = 1$, $p < .01$).

Apparently, many *S*s in Cell 1a attempted to discover a search solution but failed, and reverted to the nonsearch approach, thus inflating the time data for that cell. Furthermore, if the search criterion problem is considered to be a nonsearch problem, since many *S*s reacted to it as such, then Cells 2a and 3a represent appropriate prior experience with Cells 1a and 4a inappropriate prior experience. Of note, the former two cells took less time to solution than the latter two. Supporting evidence for these conjectures is obtained by comparing the times of those attaining search solutions in Cells 1a and 3a and comparing the times of those attaining nonsearch solutions. Since only 3 *S*s in Cell 3a attained search solutions, this

[2] The use of the category "emergent search solutions" was not necessary in this experiment since the use of a search criterion problem requiring transfer reduced the number of such emergent solutions to zero (as far as one could tell from the protocols).

data was not broken down. The 14 *S*s who attained search solutions in Cell 1a attained a solution significantly more quickly than *S*s in Cell 3a (79 seconds as compared to 97 seconds, $t = 2.18$, $p < .05$). Cell 1a *S*s who attained a nonsearch solution took significantly more time than Cell 3a *S*s (118 seconds > 97 seconds, $t = 2.59$, $p < .02$). This latter comparison supports the inference that some Cell 1a *S*s were set to look for a shortcut and looked for one; being unsuccessful they resorted to a nonsearch solution. The total of searching and adding required more time than simply adding (Cell 3a).

Furthermore, Cell 1a *S*s who attained a nonsearch solution took about the same amount of time to solution as did Cell 4a *S*s—the group which had search experience and a nonsearch criterion problem, further substantiating the inference that the former searched but failed to find the shortcut. On this basis, the search set did not appear to transfer.

EXPERIMENT III

Rationale and Procedure

Experiment III was undertaken on the assumption that some *S*s in Cell 1a in Experiment II were capable of search under transfer conditions but had abandoned search to save time, after attempting to discover a search solution on the search transfer criterion problem and not immediately finding one. In order to discover if, in fact, prior search experience led to a transferable search set, it was deemed necessary to increase the value and efficiency of the search solution. To accomplish this, Cells 1a and 3a of Experiment II were repeated using a 12 × 12 matrix of numbers as a criterion problem (rather than a 4 × 6), with the same shortcut rule as in Experiment II applying. A nonsearch criterion problem of comparable size was also generated. Since the nonsearch solution time was so greatly increased in this larger problem, the efficiency of the search solution was increased accordingly.

Forty-eight different *S*s from the same population as in the previous two experiments participated in Experiment III. Procedures in the third experiment were the same as in the previous one for Cells 1a and 3a (see Figure 2). The same prior experience problems were used in Experiment III as in the previous two experiments; only the criterion problems differed.

RESULTS

The *S*s in the two cells of Experiment III were found not to differ in time to solution (339 seconds and 340 seconds). A comparison of frequency of *S*s attaining search solutions to those attaining nonsearch solutions appears in Table 3. As the table shows, significantly more Cell 1a′ *S*s attained search solutions than did Cell 3a′ *S*s ($\chi^2 = 6.80$, $df = 1$, $p < .01$). The *S*s in Cell 1a′ attaining a search solution took significantly less time to solution than Cell 3a′ *S*s ($t = 3.36$, $df = 32$, $p < .01$), *while Cell 3a′ Ss took significantly less time to solution than Ss in Cell 1a′ who attained a nonsearch solution* ($t = 1.90$, $df = 35$, $p < .05$).[3]

TABLE 3

NUMBER OF SUBJECTS IN CELLS 1a′ AND 3a′ WHO ATTAINED SEARCH AND NONSEARCH SOLUTIONS TO THE SEARCH CRITERION PROBLEM AND MEAN TIME TO SOLUTION IN SECONDS FOR EACH SUBGROUP (EXPERIMENT III)

Solution	Cell 1a′		Cell 3a′		Total subjects
	Subjects	*M*	Subjects	*M*	
Search	11	215	2		13
Nonsearch	14	432	21	360	35
Total	25		23		48

This latter finding leads to the conclusion, as in Experiment II, that being set to search (i.e., looking for a shortcut) transfers while the use of the search set (i.e., attainment of a search solution) only partially does so. In other words, the strategy transfers to a greater extent than the skill. This inference gains greater support from the data than does an explanation in terms of the relative efficiencies of search and nonsearch solutions.

DISCUSSION

From the findings of the three experiments, it was possible to make an inference that was both important and unexpected, namely that *the strategy of*

[3] Two *S*s in Cell 3a′ attained a search solution to the criterion problem (see Table 3). These two *S*s completed the criterion problem in half as much time as the mean for their cell. Had these two *S*s been removed from the Cell 3a′ data, the difference between the times of nonsearchers in Cell 1a′ and the remaining *S*s in Cell 3a′ (all nonsearchers) would have been even more marked than the difference shown.

search could be made more readily to transfer than the skill of search, as the result of limited prior experience.

In the initial formulation, search strategy and search skill were incorporated into the single concept of search set with the expectation that such sets could be induced by providing appropriate experiences. This expectation appeared to hold in the first experiment where the search criterion problem and search experience problems were quite similar (and furthermore, the shortcut in the search criterion problem was apparently "easy" to find). Analysis of protocols showed that the situation was *not* one where *S*s *either* looked for *or* did not look for a shortcut. Among *S*s receiving nonsearch experience, a shortcut solution to the search criterion problem was apparently "stumbled upon" (i.e., *emerged*) in the majority of cases. This, coupled with the significant main effect of criterion problem, indicates that the ease of finding the shortcut (a problem characteristic) was an important determinant of the solution strategy adopted by *S*, even when contrary to the practice sequences provided by the experimenters. (This was an unintended outcome.) However, the greatest influence on strategy adoption was the joint effect of search experience and the visibility of the shortcut in the search criterion problem.

The visibility of the shortcut was drastically reduced in the criterion problem used in the second experiment. This is indicated in part by the absence of a main effect of criterion problem in the second experiment and the fact that only 28% of *S*s exposed to search experience and 6% of *S*s not so exposed recognized the shortcut solution (as opposed to 89% and 69%, respectively, in the first experiment). Thus, the experience effect and problem effect are separated in the second experiment by virtue of the elimination of the latter.

The experience effect in the second experiment is clearly not the simple one predicted in the second hypothesis. The search experience sequence appeared to have both enhancing and debilitating effects on performance by *S*s on the search transfer criterion problem. The nonsearch experience sequence on the other hand had no differential effect on search criterion problem versus nonsearch criterion problem performance (compare Cells 2a and 3a). The result was a significant interaction between prior experience and criterion problem and prompted internal analyses. By splitting the group that received search experiences prior to the search criterion problem into those who attained a search solution and those who did not, it was possible to pinpoint the differential effect of search experience on strategy and skill. Specifically, *S*s attaining a nonsearch solution on the search criterion after search experience as compared to after nonsearch experience required more time to solution. Apparently such *S*s spent time looking for a shortcut as a result of their search experiences but had not had sufficient experience to develop the level of skill required to find a shortcut. Eventually they "gave up" and reverted to a nonsearch solution. Findings of the third experiment supported this interpretation.

The consistent findings of this study lead to the conclusion that search skill has quite limited transfer possibilities, certainly as compared to search strategy (i.e., searching as compared to finding), and stimulates the recommendation that the conceptual distinction between these two phenomena be retained, and that they be labeled differently. (The initial position in this paper was to combine them, which does not now appear to be conceptually sound.)[4] It would seem reasonable to retain the term "search set" to refer to the strategy of search and use some other term, perhaps "learning set," to refer to the skill in finding a shortcut solution. A specific experience or set of experiences does not appear to affect both the strategy of search and the skill involved in a particular kind of searching to the same

[4] It must be emphasized that the main finding of this experiment was serendipitous insofar as the experimenters let the data lead them to the unavoidable conclusions. When breaking new ground, this appears to be a useful research strategy.

degree. Thus, these two phenomena are different in a practical as well as a conceptual sense.

One implication of this finding is that limited educational exposure to elegant thinking and problem-solving approaches may induce students to adopt the strategy to search when confronted by transfer situations but leave them lacking the skill to successfully apply this strategy. The result would be performance inferior to the inelegant technique, and perhaps frustration. Based on this implication, one must take care to provide a level of skill commensurate with a student's commitment to a strategy in order that he can use this strategy effectively, if at all. More extensive experience sequences than those provided in this experiment would be needed.

It was concluded from the first experiment that suggestion has relatively little effect on performance. Certainly suggestion provides *S* with no environmental evidence for the efficacy of the strategy.

REFERENCES

BRUNER, J. S., GOODNOW, J. J., & AUSTIN, G. A. *A study of thinking*. New York: Wiley, 1956.

GAGNÉ, R. M., & BROWN, L. T. Some factors in the programming of conceptual learning. *Journal of Experimental Psychology,* 1961, **62,** 313–321.

GAGNÉ, R. M., MAYOR, J. R., GARSTENS, H. L., & PARADISE, N. E. Factors in acquiring knowledge of a mathematical task. *Psychological Monographs,* 1962, **76**(7, Whole No. 526).

GAGNÉ, R. M., & PARADISE, N. E. Abilities and learning sets in knowledge acquisition. *Psychological Monographs,* 1961, **75**(14, Whole No. 518).

GOODNOW, J. J. Determinants of choice distributions in two-choice probability situations. *American Journal of Psychology,* 1955, **68,** 106–116.

GOODNOW, J. J., & PETTIGREW, T. F. Effect of prior patterns of experience upon strategies and learning sets. *Journal of Experimental Psychology,* 1955, **49,** 381–389.

HARLOW, H. F. The formation of learning sets. *Psychological Review,* 1949, **56,** 51–65.

HARLOW, H. F. Learning set and error factor theory. In Sigmund Koch (Ed.), *Psychology: A study of a science*. Vol. 2. New York: McGraw-Hill, 1959. Pp. 492–537.

KATONA, G. *Organizing and memorizing*. New York: Columbia University Press, 1940.

LUCHINS, A. S. Mechanization in problem solving. The effect of *Einstellung*. *Psychological Monographs,* 1942, **54**(6, Whole No. 248).

PEPITONE, A., & REICHLING, G. Group cohesiveness and the expression of hostility. *Human Relations,* 1955, **8,** 327–337.

WINER, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

(Received December 20, 1966)

Journal of Educational Psychology
1968, Vol. 59, No. 2, 69–73

DELAYED INFORMATION FEEDBACK, FEEDBACK CUES, RETENTION SET, AND DELAYED RETENTION

JULIUS M. SASSENRATH AND GEORGE D. YONGE
University of California, Davis

8 groups of about 20 undergraduates each were presented with 60 factual multiple-choice items, answered each question, and either received feedback immediately or 24 hr. later. For 4 groups feedback included the stem and 4 alternatives to each question. 4 groups received only the 4 alternatives. Just prior to feedback 4 groups were given a retention set and 4 groups were not. The design of the experiment was 2 (feedback) × 2 (stem, no stem) × 2 (set, no set). A 60-item immediate and 5-day delayed retention test was administered. On immediate retention only the effect of stem, no stem was significant. On the delayed retention test the groups receiving delayed feedback, the stem of the question, and the retention set performed reliably higher than their counterparts.

The idea that learning is improved when a reinforcer or some information feedback (IF) promptly follows one's behavior is found in several prominent theories of learning (Hull, 1952; Skinner, 1957; Spence, 1956). Most textbooks in psychology and educational psychology mention the superiority of immediate information feedback (IIF) as a principle of human learning and point out its presumed importance in educational practice. The advent of teaching machines has increased the acceptance of this principle among educators and most psychologists. However, the application of IF is not limited in education to teaching machines. Thus, telling a student his answer is "right," or quickly handing back examinations with corrections and comments, are examples of IF. Sassenrath and Garverick (1965) have recently found that different types of IF on midterm examinations, 2 days after taking the examination, have different effects on retention and transfer of learning.

Evidence for the effects of reinforcement or IF on learning comes from three general kinds of studies: (*a*) animal experiments; (*b*) experiments of human motor skills; and (*c*) experiments employing verbal materials. In the area of animal learning, Renner (1964) reviewed the literature on delay of reinforcement. The evidence indicates that learning efficiency decreases the longer the delay of reinforcement. On this point there is certainly a big question regarding the logic of inferring from a rat learning to press a bar or a pigeon learning to peck at a disk to a child learning to speak his native language or an adult learning matrix algebra. The implication is that the principle of reinforcement may not have the generality it is commonly assumed to have. Furthermore, the learning of factual multiple-choice materials, which will be reported in the present study, may not have much to do with learning a language or mathematics either.

Evidence from studies on human motor skills calls into question the superiority of IIF on learning. Of 14 studies dealing with this problem, 11 show no significant difference in learning efficiency, one favors IIF, and two favor delayed information feedback (DIF) (Brackbill, Wagner, & Wilson, 1964). In experiments dealing with verbal, concept, or discrimination learning materials contradictory results have been obtained. Saltzman (1951) and Bourne (1957) report learning impairment associated with increased DIF. On the other hand, Brackbill and associates (Brackbill, Boblitt, Davlin, & Wagner, 1963; Brackbill, Bravos, & Starr, 1962; Brackbill, Isaacs, & Smelkinson, 1962; Brackbill & Kappy, 1962; Brackbill, Wagner, & Wilson, 1964) have generally found no difference in rate of learning between IIF and DIF. These latter findings are important since most of the previous studies

with humans and animals, and most the theories of learning, have taken the position that DIF or delayed reinforcement either had no effect or impaired rate of learning.

Equally important for psychological theory are the findings on children by Brackbill and on adults by Sturgis (1966) and Sturgis and Crawford (1963) that DIF produces higher *retention* scores than does IIF under some conditions. For education these results also are important. Why teach or learn something if it is not remembered after some brief period of time? One goal of education is not just to have students learn but to have them remember what they have learned so that they may use or transfer what they remember to other learning situations or so they can use their knowledges and skills to solve problems.

One of the possible theoretical explanations for the beneficial effect of DIF on retention is the hypothesis that during the delay interval the learner can use response-produced, external, or verbal cues to help mediate the DIF period (Brackbill & Kappy, 1962). With verbal materials and highly verbal older children or college students, it would appear that the use of verbal cues might be extremely important in regulating one's behavior with respect to memory of past events and the anticipation of future events. This may be one reason why delayed reinforcement with animals impairs learning (Renner, 1964), whereas with children and adults DIF did not impair learning and did facilitate retention.

In addition, in animal studies of delayed reinforcement, the animal has to remember his response over the delay interval and is then given only the reinforcement without being again presented with the original task and alternatives. In the studies with humans, the subject (*S*) also has to remember his response over the delay period but during IF he is usually presented with the question or task as well as the alternatives he should or should not have made. Thus, the partial or complete re-presentation of the initial task (or question and alternative) appears to be an important issue, and one which will be inquired into in the present study.

Related to the notion that external or verbal cues between response and IF facilitates retention is the fact that a set to remember (instructional cues) when given to students *before* learning (feedback) enhances retention (Karen, 1956). On the other hand, a set to remember given students *after* learning (feedback) does not enhance retention (Ausubel, Schpoont, & Cukier, 1957). Perhaps this is because the way the material is learned and remembered under a retention set is different than under no retention set—even though the same amount of material may be learned.

In her research with college students, Sturgis (1966) found that DIF produced superior delayed retention without influencing immediate retention, but only when students were given complete feedback cues. However, in her study and in a previous one (Sturgis & Crawford, 1963), Sturgis confounded IIF per question with 24-hour DIF per list. Therefore, the purpose of the present experiment was to correct this methodological difficulty and to ascertain the effect on immediate and delayed retention of (*a*) DIF, (*b*) the amount of IF cues, and (*c*) a set to remember.

METHOD

Materials and Procedure

Sixty, four-alternative, factual, multiple-choice questions generally found in courses in introductory psychology constituted the learning and retention materials. Thirty-eight of these questions were used by Sturgis (1966). Each question was mimeographed on an 8.5 × 3-inch piece of paper and the 60 questions were stapled together into a packet. Each student was given a packet and an IBM answer sheet for the initial presentation of the list of questions. Students were given 15 seconds to read and mark an answer to each of the 60 questions. After the initial presentation of the list, the packet and the answer sheets were collected. The group which received IIF was immediately given another packet of 60 questions. Half of the students received a packet with the stem cues of the question and the four alternatives with the correct alternative underlined. The other half of the IIF group received a packet of the 60 items with the four alternatives and the correct alternatives underlined, but there was no stem

cue presented. In both cases students were given 10 seconds to read each item during the IIF period. About half of the students receiving stem cues and no stem cues were told just before the IIF that they should try hard to learn and remember the correct answer to each question since there would be a retention test later. This is called a retention set. The other half of the students received no retention set.

Half of the students who took the initial presentation of the 60 questions received 24-hour DIF. As was done with the IIF students, half of the DIF students received the stem cues of each question and the four alternatives with the correct one underlined. The other half did not see the stem cues. Both groups received a 10-second IF exposure on each item. Again, as was done with the IIF students, half the students under the DIF condition that received the stem and no stem were given the retention set immediately before the DIF. The other half had no retention set.

Immediately after completing either the IIF or the DIF, the packet of materials was collected and a mimeographed retention test of the 60 multiple-choice questions and an IBM answer sheet was passed out to each student. The questions on the immediate retention test were in a different order than on the initial presentation. Students could answer the questions on the retention test at their own rate, and most students finished in about 10 minutes. This is called the immediate retention test. Five days later, students were administered the same retention test and worked through it at their own rate. The questions were arranged in a different order again. Most students finished this delayed retention test in about 12 minutes.

Subjects and Design

The students were 163 upperclassmen enrolled in four sections in a course in introduction to educational psychology. The experiment was a 2 (IIF versus DIF) × 2 (stem cues versus no stem cues) × 2 (retention set versus no retention set) design with about 20 students in each of the eight groups. The experimental treatments were assigned to the sections at random. Conflicts in class schedules precluded assigning students to sections at random.

RESULTS

Table 1 presents the mean scores on the initial, immediate retention, and delayed retention tests for students in each of the three experimental conditions. A 2 × 2 × 2 analysis of variance (with a correction for unequal *S*s in each group) of the *initial* test scores indicates that there was no significant difference among the groups due to IF, feedback cues, or retention set. In addition, none of the interactions was significant. Thus, the students in the subsequent experimental conditions were on the average about equal in terms of their scores on the initial test.

A three-way analysis of variance (with correction for unequal N) of the *immediate* retention test showed that there were no significant differences due to IF or retention set, but the difference due to feedback cues was significant, $F = 38.53$, $df = 1/156$, $p < .001$. Apparently, whether IF is immediate or delayed 24 hours after the initial presentation has no immediate effect upon retention. Also being told to remember what one is about to learn has no effect on immediate retention. However, when IF included the stem of the question there was

TABLE 1

MEANS AND STANDARD DEVIATIONS ON THE INITIAL, IMMEDIATE RETENTION, AND DELAYED RETENTION TESTS

Experimental conditions	N	Initial M	Initial SD	Immediate M	Immediate SD	Delayed M	Delayed SD
Information feedback							
Immediate	81	28.0	4.7	55.4	3.0	50.8	3.5
Delayed	82	27.8	4.2	56.1	3.0	52.9	4.4
Feedback cues							
Stem cues	81	28.0	4.5	56.7	2.5	54.5	3.4
No stem cues	82	27.8	4.3	53.8	3.5	49.2	4.5
Retention set							
Yes	87	28.3	4.8	56.0	2.5	52.6	3.3
No	76	27.7	4.1	55.4	3.5	50.9	4.6

a greater effect upon immediate retention than when IF did not include the stem of the question. Notice, however, that *S*s who received no stem cues still learned a great deal as measured from their initial performance. For that matter, notice that the increase in scores from the initial to the immediate retention test due to learning is about 100%. None of the interactions on the immediate retention test was significant.

On the *delayed* retention test, the analysis of variance (with correction for unequal *N*) indicated that the differences were significant for IF, $F = 5.06$, $df = 1/156$, $p < .05$, for retention set, $F = 7.69$, $df = 1/156$, $p < .01$, and for feedback cues, $F = 45.94$, $df = 1/156$, $p < .001$. None of the interactions was significant. Thus, as can be seen in Table 1, *S*s who received DIF performed reliably higher on delayed retention than *S*s who received IIF. Also *S*s who received stem cues during feedback did better than *S*s who received no stem cues. And *S*s who received a retention set performed higher on delayed retention than *S*s who did not receive a retention set.

DISCUSSION

The results of the present study indicate that DIF and IIF during learning produce no differential effect on a retention test administered immediately after learning. On the other hand, 5 days later on the delayed retention test, there is a small but significant difference favoring *S*s who received DIF. Both Sturgis (1966) and Sturgis and Crawford (1963), also working with college students and employing similar materials, have found similar results. This was true in spite of the fact that Sturgis confounded IIF per item with DIF per list of items. The only difference was that Sturgis (1966) found an interaction in that DIF was superior only when *S*s received the right and wrong alternatives along with the stem of the question during IF. The *S*s in the study by Sturgis that received DIF with the question and *only* the correct alternative did not perform reliably higher than *S*s that received IIF. In their studies with third graders, Brackbill and associates (Brackbill, Boblitt, Davlin, & Wagner, 1963; Brackbill, Bravos, & Starr, 1962; Brackbill, Isaacs, & Smelkinson, 1962; Brackbill & Kappy, 1962; Brackbill, Wagner, & Wilson, 1964) also have generally found that *S*s receiving DIF, even of only 10 seconds, performed higher on relearning than *S*s that received IIF, and also generally found that DIF did not impair initial learning. Thus, there is mounting evidence that DIF does not retard learning and may enhance delayed retention. If so, these results have considerable implications for learning theory, programmed instruction, and classroom teaching.

How does one explain the effect of DIF on retention? One hypothesis is that with DIF *S*s can make use of more cues from the initial task or question before DIF is presented (Brackbill & Kappy, 1962; Renner, 1964). For human *S*s with a verbal repertoire, these cues are largely verbal and assist *S*s in mediating the DIF period and/or the period between learning and retention. However, if this hypothesis were tenable, why is it that in the present experiment and in the studies by Sturgis the effect of DIF was not found on the immediate retention test but only appeared on the delayed retention test? In other words, whatever brings about the effect of DIF on delayed retention does not occur *solely* before, during, or immediately after feedback (learning). Otherwise, the effect should occur on the immediate retention test. It would appear that there might be some interaction between what happens during the DIF process and the interval prior to or during the administration of the *delayed* retention test. It could be speculated that during the DIF period, when a verbal *S* responds initially to a question, certain response-produced verbal cues may be implicitly or covertly rehearsed. This initial rehearsal may not have any beneficial effect on immediate retention, but may bring about a second covert or overt verbal rehearsal of the question during the delayed retention interval or at the time of the delayed retention test. Thus, the verbal cues during the feedback process may influence the use of

verbal cues during the retention process. An experiment is now underway to determine if various kinds of cues presented during DIF or IIF have any effect on immediate and delayed retention.

Ausubel, Schpoont, and Cukier (1957) argue that set to remember does not produce superior retention per se but rather results in superior learning which, in turn, produces greater retention. The results of the present study, however, indicate that a set to remember produces superior results on the delayed retention test as compared with a no-set condition, but no difference on the immediate retention test. That is, there is no evidence for differential learning even though differential retention was evident. This would suggest that the meaning of the task, rather than degree of learning, was modified by set instructions.

In the present study it was also found that *S*s receiving cues during IF, which included the stem of the question and the alternatives, performed reliably higher on both the immediate and delayed retention tests than did *S*s receiving no stem cues but only the alternatives. In other words, *S* had to remember his response over the period prior to DIF but was then given either full or partial IF which included the question or task as well as the alternatives they should and should not have made. In animal studies of delay reinforcement *S* also has to remember his response over the delay interval but is then only given reinforcement without also being presented with the initial task and alternatives. Under these conditions animals receiving a long (over 10 seconds) delay in reinforcement show little if any learning or retention. On the other hand, human *S*s under 24-hour DIF condition who received only the alternatives without the task or question still show considerable learning and retention. However, they do not show as much retention as *S*s presented with the task and alternatives at DIF.

REFERENCES

AUSUBEL, D. P., SCHPOONT, S. H., & CUKIER, L. The influence of intention on the retention of school materials. *Journal of Educational Psychology*, 1957, **48**, 89–92.

BOURNE, L. E. Effects of delay of information feedback and task difficulty on the identification of concepts. *Journal of Experimental Psychology*, 1957, **54**, 201–207.

BRACKBILL, Y., BOBLITT, W. E., DAVLIN, D., & WAGNER, J. E. Amplitude of response and the delay-retention effect. *Journal of Experimental Psychology*, 1963, **66**, 57–64.

BRACKBILL, Y., BRAVOS, A., & STARR, R. H. Delay improved retention of a difficult task. *Journal of Comparative and Physiological Psychology*, 1962, **55**, 947–952.

BRACKBILL, Y., ISAACS, R. B., & SMELKINSON, N. Delay of reinforcement and the retention of unfamiliar meaningless material. *Psychological Reports*, 1962, **11**, 553–554.

BRACKBILL, Y. & KAPPY, M. S. Delay of reinforcement and retention. *Journal of Comparative and Physiological Psychology*, 1962, **55**, 14–18.

BRACKBILL, Y., WAGNER, J. E., & WILSON, D. Feedback delay and the teaching machine. *Psychology in the Schools*, 1964, **1**, 148–150.

HULL, C. L. *A behavior system.* New Haven: Yale University Press, 1952.

RENNER, K. E. Delay of reinforcement: A historical review. *Psychological Bulletin*, 1964, **61**, 341–361.

SALTZMAN, I. J. Delay of reward and human verbal learning. *Journal of Experimental Psychology*, 1951, **41**, 437–439.

SASSENRATH, J. M., & GARVERICK, C. M. Effects of differential feedback from examinations on retention and transfer. *Journal of Educational Psychology*, 1965, **56**, 259–263.

SKINNER, B. F. *Verbal behavior.* New York: Appleton-Century-Crofts, 1957.

SPENCE, K. W. *Behavior theory and conditioning.* New Haven: Yale University Press, 1956.

STURGIS, P. T. Delay-improved retention as a function of number of alternatives presented in knowledge of results. Paper presented at the meeting of the Western Psychological Association, Long Beach, April 1966.

STURGIS, P. T., & CRAWFORD, J. J. The effect of delay of reinforcement on learning factual material. Paper presented at the meeting of the Western Psychological Association, Santa Monica, April 1963.

(Received March 1, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 2, 74–81

BEHAVIORAL CORRELATES OF ACADEMIC ACHIEVEMENT

II. PURSUIT OF INDIVIDUAL VERSUS GROUP GOALS IN A DECISION-MAKING TASK[1]

ROBERT S. WYER, JR.[2]

University of Illinois, Chicago Circle

College freshmen, representatives of 4 combinations of academic aptitude (college entrance examination score) and academic performance (1st-term grade-point average), participated in 2-person groups in a decision-making task in which their choices either would increase the likelihood of individual goal attainment at the expense of group goal attainment, or would increase the likelihood of attaining group goals at the sacrifice of individual goals. As expected, there was a general tendency to increase the frequency of group-oriented choices when group goals were greatly affected. When choices had relatively little effect upon the likelihood of attaining group goals, the frequency of group-oriented choices (choices that would increase the likelihood of group goal attainment) was related positively to academic performance among both males and females. When decisions had relatively great effect upon group goals, this relationship occurred only among Ss of high academic aptitude. Results raised questions concerning the general validity of the assumption that socially oriented behavior tendencies are detrimental to academic effectiveness.

Research on the motivational correlates of academic achievement (cf. Pierce & Bowman, 1965; Todd, Terrell, & Frank, 1962) has emphasized two factors: The desire for personal achievement and the desire for effective social relations. Wyer and Terrell (1965) found significant sex differences in the relationships between academic performance and the acknowledged desire for recognition in both academic and social areas. In general, however, studies of motivational factors associated with academic effectiveness have yielded fairly inconclusive results (Lavin, 1965, pp. 74–79).

The effect of social motivation upon academic achievement has been especially unclear. It is often assumed that a social orientation is detrimental to the pursuit of academic goals. This presupposes an incompatibility between socially directed behavior and academic goal attainment. However, the willingness to cooperate with other persons in achievement-related activity may often lead to more effective goal seeking; if manifested in academic areas (e.g., through participation in informal group discussions, giving and receiving assistance in problem solving, etc.), it may increase academic effectiveness. The study reported here, one of a series of studies of nonintellective factors associated with academic success in college (Wyer, 1965, 1967; Wyer & Terrell, 1965; Wyer, Weatherley, & Terrell, 1965), was concerned with this issue. A situation was constructed in which subjects (*S*s) were faced with a decision either to seek goals independently or to seek them in cooperation with other persons. The tendency to respond cooperatively in this situation was analyzed as a function of academic aptitude and performance. These analyses were expected to provide information on the facilitative or detrimental effects of socially oriented behavioral tendencies upon academic achievement.

Representatives of four combinations of aptitude (college entrance examination score) and performance (grade-point average) were divided into two-person groups and asked to participate in a decision-

[1] This research was supported by the United States Office of Education, Cooperative Research Project Contract 1139A, under the direction of Glenn Terrell.

[2] Thanks are extended to Thomas Gronek, Douglas Johnson, and Patricia Creedon for their assistance in various phases of this research.

making task. They were told that prizes would be given both for individual performance and for team performance. The task was similar to that described by Deutsch (1962). That is, *S* was given two choices (X_1 and X_2) and was told that his partner would also have two choices (Y_1 and Y_2). The outcome in terms of the number of points received was ostensibly determined by the conjunction of the two players' choices. As an example, consider the following matrix:

		Player Y	
		Y_1	Y_2
Player X	X_2	7,7	2,8
	X_1	8,2	3,5

For each combination of choices the first number indicates the outcome for X and the second number the outcome for Y. For instance, if X_2 and Y_2 were selected, X would receive 2 points and Y would receive 8 points; the number of points that players would receive as a team is the sum of these, or 10 points.

Eight different matrices similar to the one above were constructed and presented to *S*s sequentially. The *S*s were told that both the total number of points each player had won as an individual and the total number of points each team had won would be determined after the experiment, and that monetary rewards would be given both to persons who had accumulated the most points as individual players and to groups who had accumulated the most points as a team. The possible outcomes of various combinations of responses were made known to each *S*; however, participants were given no knowledge of their partner's choices while playing the game.

Several factors may affect decisions in situations such as the one described: One's expectancy for what his partner will do, the effect of one's choice upon both his own and his partner's outcomes, the relative value attached to team and individual goals, etc. The dynamics underlying decision-making processes in these situations are complex and have been studied extensively by Deutsch and his colleagues (cf. Deutsch, 1962). The relative contribution of these factors can be manipulated fairly successfully by varying the relative magnitude of outcomes in the decision matrix. In the present study, a set of matrices was constructed that would allow inferences to be made concerning each *S*'s relative preference for seeking individual goals (based upon the number of points he accumulated as an individual player) and team goals (based upon the combined number of points accumulated by himself and his partner). For example, in the matrix presented above, the selection of X_2 would be assumed to indicate a team orientation, while the selection of X_1 would be assumed to indicate a decision to seek individual goals.

A preference for individual rather than team goals may have several determinants. For example, it may indicate a desire to receive recognition as an individual rather than as a member of a winning group; or, it could be one manifestation of a general tendency to be socially aloof and autonomous. Alternatively, a high team orientation could indicate a desire to please or ingratiate one's partner. More simply, it may indicate that *S* has been reinforced more frequently in the past for cooperative activity than for independant activity. Regardless of these factors, the relationship between academic performance and the tendency to seek goals cooperatively rather than individually was expected to have implications for the effect of socially directed behavior upon academic effectiveness. For example, if academic achievement is facilitated by a general tendency to persevere in achievement-related activity independently of other persons, a negative relationship between academic performance and the frequency of team-oriented responses would be expected. On the other hand, if a general tendency to cooperate with other persons in achievement-related activity increases academic effectiveness, the relationship between team orientation and academic performance may be positive.

METHOD

Selection of Subjects

Participants in the study were selected from freshman liberal arts and science students at an urban midwestern university for whom measures of aptitude and performance were available. Performance (Per) was measured by first-quarter grade-point average. Aptitude (Apt) was measured by composite score on the American College Testing Service college entrance examination. Defined in this way, aptitude was not considered to be an index of intelligence, but rather was assumed to reflect the degree to which a student has mastered general intellectual skills prior to entering college, and therefore a predictor of the level of performance he should obtain with an average amount of effort.

Measures of Apt and Per were each converted to z-scores and *S*s placed into four categories according to the following criteria:

1. High Apt, high Per—z-scores of greater than .50 for both Apt and Per, and an absolute difference between Apt and Per z-scores of less than .30.
2. Low Apt, low Per—z-scores of less than −.50 for both Apt and Per, and an absolute difference between Apt and Per z-scores of less than .30.
3. High Apt, low Per—An Apt z-score of greater than .50, a Per z-score of less than −.50, and a difference between Apt and Per z-scores of greater than 1.30.
4. Low Apt, high Per—An Apt z-score of less than −.50, a Per z-score of greater than .50, and a difference between Per and Apt z-scores of greater than 1.30.

Sixteen males and 16 females were selected at each of the four combinations of Apt and Per and recruited for the experiment. They were each paid $1 for their services.

Construction of Matrices

Two criteria were used in preparing decision-making matrices. First, *S*'s choice should unequivocally reflect a decision to pursue either a team goal or an individual goal; to meet this criterion, any alternative that maximized the number of points that the group would receive minimized the number of points *S* would receive as an individual player, and vice versa. Second, *S*'s decision should depend minimally upon the response he expects his partner to make. This was done either by making the partner's choice clear, or by insuring that an *S*'s choice could be interpreted similarly regardless of the partner's choice. The eight matrices selected, in the order of their presentation to *S*s, are shown in Table 1. (In each case, assume that *S* is Player X.)

The effect of *S*'s choice upon team outcomes relative to the effect of his choice upon individual outcomes varied over matrices. To indicate the extent of this variation, the effect of *S*'s choice upon the number of points he received as an individual player (I) was determined by subtracting the number of points he would receive by making one selection from the number he would receive by making the other choice. The choice that Player Y would be expected to make in responding to Matrices 2 and 4 were clear; for these, the number of points X would receive given this choice (Y_1) was used in the calculations. For the other matrices, in which Y might be expected to make either response, the mean number of points X would receive as an individual player was averaged over Y's alternatives. The effect of *S*'s choice upon team outcomes (T) was determined similarly. The relative effect of *S*'s choice on team outcomes relative to individual outcomes (D_{TI}) was then calculated for each matrix by subtracting I from T. Values of I, T, and D_{TI} for each matrix are shown in Table 1.

It could be argued that *S*s who adopt a team strategy in responding to these matrices do so not because they have interest in cooperative pursuit of goals per se, but because they believe this strategy to maximize the probability of receiving a reward, independent of its group or individual nature. This, however, seems unlikely. Which strategy is really optimal is difficult to discern. For example, on Matrix 1 it appears that if X

TABLE 1

SUMMARY OF MATRICES PRESENTED TO SUBJECTS AND THE EFFECT OF CHOICES ON OUTCOMES

Matrix	Subject's choice	Partner's choices		Average effect of choice		
		Y_1	Y_2	On team outcome (T)	On individual outcome (I)	D_{TI}
1	X_2	7,7	2,8	3	1	2
	X_1	8,2	3,5			
2	X_2	5,4	4,2	2[a]	3[a]	−1
	X_1	2,9	5,3			
3	X_2	5,5	6,6	1	1.5	−.5
	X_1	6,3	8,3			
4	X_2	5,5	6,4	2[a]	2[a]	0
	X_1	3,9	4,8			
5	X_2	5,5	0,7	4	1.5	2.5
	X_1	7,0	1,1			
6	X_2	7,4	5,3	2	1	1
	X_1	6,6	4,7			
7	X_2	9,2	7,3	4	1	3
	X_1	8,8	6,7			
8	X_2	4,0	5,1	7	1	6
	X_1	4,9	3,8			

[a] Calculation based upon assumption that the partner would select Y_1.

played individually he would have an excellent chance of attaining a group goal as well as individual goal should Y use a group strategy, and would not lose much (relative to Y) should Y play individually. On subsequent matrices, however, a group strategy often appears optimal. Furthermore, it should be noted that if X plays for a team goal he increases Y's chances of receiving *both* a group goal and an individual goal; a decision to do so therefore seems highly cooperative.

If the total points attainable by X and Y under the four possible combinations of strategies used by the two players are calculated, it can be seen that the optimal strategy for Y would be to play individually, because he will maximize individual outcomes and, provided X adopts a group strategy, would also have high team outcomes. X, if he were to realize this, would be confronted with the decision to take a chance on winning a group goal by obtaining high (87–92) but not optimal (100) payoff, or to try for a high (38–43) but not optimal (55) individual score. If he decides the former, he decreases his chances for an individual reward (since Y will be above him) but remains high in the running for a group goal regardless of Y's behavior. On the other hand, if he decides to seek an individual goal, he insures that he is higher than at least one other competitor (Y). The difficulty of determining which orientation will maximize the likelihood of receiving a reward seems to justify the assumption that differences in orientation primarily reflect differences in the relative preference for pursuing goals cooperatively versus individually. (In this regard, *Ss* typically took less than 10 minutes to complete the task; since the time required to determine the optimal strategy to use across matrices would be substantial, this fact also argues against the alternative interpretation in question.)

Administration Procedure

Four or five pairs of *Ss* were administered the task simultaneously. In each case, partners were of the same sex. In four instances in which an *S* did not show up for the experiment at the scheduled time, the "odd" *S* was informed that a person had been left over during a previous session and was told to assume that this person was his partner.

Partners were placed beside one another at long tables, far enough apart so that they could not see each other's work. They were given booklets containing one sample matrix and the eight test matrices described above. One member of each team was designated Player A and the other Player B. The *Ss* were led to believe that the matrices they were presented indicated both their own and their partner's outcomes. In fact, the forms distributed to all *Ss* were identical except for the sample matrix and their designation as either Player A or Player B. (In Table 1, Player X was always the *S*, regardless of whether he was formally assigned to be A or B.)

To explain the task, *Ss* designated as Player A were presented the sample matrix below:

		Player B	
		B_1	B_2
Player A	A_2	5,6	3,6
	A_1	5,3	4,3

The *Ss* designated as B were presented a similar matrix, rearranged so that B's outcomes were listed first. All *Ss* were read the following instructions:

> We are interested in determining how persons behave when their behavior affects not only their own goals but the goals attained by others. I am going to ask you to play a game with the person next to you. One player, labeled A, will be able to choose either A_1 or A_2; the second player, B, will have to choose between alternatives marked B_1 and B_2. On any given trial, each player will be awarded a certain number of points. The number of points he wins will depend not only upon his own choice but also on the choice of his partner.

The *Ss* were then referred to the sample matrix on the first page of their booklet and the outcomes of each combination of choices were explained. The instructions then continued:

> On the form I have passed out there are eight tables similar to this one. Below each table, the possible combinations of choices and outcomes are written down. Both you and your partner will have three trials in each game. At no time, however, will you know how your partner has moved before making your own decision. In planning your move, you will therefore have to guess how he is likely to respond. The number of points each player wins will be determined after the game by comparing the choices each player has made on corresponding trials.

To provide an incentive to perform well on the task, and also to make clear to *Ss* that they could work either for individual goals or for team goals, the following additional instructions were read:

> Each player will receive a score based upon the total number of points he has accumulated. On the other hand, each team will also receive a score based upon the total number of points both partners together have accumulated. To make the game interesting, we will award a prize of $2.50 to each of the ten individuals in the entire experiment who accumulate the greatest number of points for themselves, and a prize of $2.50 to each player on the five teams who have accumulated the greatest number of points as a team. There are 128 persons competing on 64 teams; your chance of winning either a team prize or an individual prize is therefore about one out of six.

Each *S* was asked to make 3 responses to each matrix. For each *S*, the number of team-oriented responses, or the number of responses that would maximize the total number of points awarded to

S and his partner, was determined for each matrix and was analyzed as a function of aptitude, performance, and sex.

RESULTS

Four *S*s who were selected for the study were unable to participate; three more *S*s did not understand the instructions and recorded their answers incorrectly. To obtain proportional cell frequencies necessary for analysis of variance, seven more *S*s were eliminated at random from various cells. The final sample consisted of 15 *S*s of high ability and 13 *S*s of low ability at each combination of sex and performance.

Some indication of the effectiveness of the experimental procedure in producing team-oriented and individual-oriented behavior could be obtained by comparing the frequency of team-oriented responses to a particular matrix with the magnitude of the effect of these responses on team outcomes. If *S*s understood the experimental task and the consequences of their choices, they should generally make more team-oriented responses when the magnitude of team outcomes was relatively more affected by their choices. This appeared to be the case. The correlation between the mean number of team-oriented responses to each matrix and D_{TI} was .77 ($N = 8$, $p < .025$). The correlation calculated for each *S* and then averaged over *S*s was lower ($M = .21$, $N = 112$), but also was in the direction expected.

An extension of a Lindquist Type III analysis of variance was performed on the number of team-oriented responses as a function of sex, aptitude, performance and type of matrix. (Data relevant to the analysis are shown in Table 2.) This analysis yielded significant main effects of performance ($F = 5.07$; $df = 1/104$; $p < .05$; $MS_e = 2.81$) and matrix ($F = 10.08$; $df = 7/728$; $p < .01$; $MS_e = .445$) and a significant interaction of sex, performance, and matrix ($F = 2.10$; $df = 7/728$; $p < .05$).

The main effect of matrix type indicated that team orientation increased with D_{TI}, as also noted in correlation analyses. Low performers were less team-oriented ($M = 1.74$) than high performers ($M = 2.15$). To explore the contingencies indicated by the significant interaction, supplementary analyses were performed involving (*a*) only the four matrices for which D_{TI} was

TABLE 2

MEAN NUMBER OF TEAM-ORIENTED CHOICES AS A FUNCTION OF APTITUDE (APT), PERFORMANCE (PER), SEX, AND TYPE OF MATRIX

Group	Number of matrix								*M*
	2	3	4	6	1	5	7	8	
	$D_{TI} = -1$	$D_{TI} = -.5$	$D_{TI} = 0$	$D_{TI} = 1$	$D_{TI} = 2$	$D_{TI} = 2.5$	$D_{TI} = 3$	$D_{TI} = 6$	
Low Apt, low Per									
Males	1.46	1.53	1.00	1.54	1.77	2.15	2.23	2.07	1.72
Females	1.46	1.77	1.77	2.00	1.38	2.08	1.62	2.31	1.80
M	1.46	1.65	1.38	1.77	1.58	2.16	1.92	2.19	1.76
High Apt, low Per									
Males	1.33	1.67	1.13	1.80	1.40	2.00	1.67	1.60	1.58
Females	1.60	1.93	1.60	1.93	1.60	2.00	2.13	2.13	1.86
M	1.47	1.80	1.37	1.87	1.50	2.00	1.90	1.87	1.72
Low Apt, high Per									
Males	1.85	1.85	1.54	1.61	1.69	1.54	1.69	1.69	1.68
Females	1.77	1.85	1.69	2.08	1.92	2.08	2.23	2.15	1.97
M	1.81	1.85	1.62	1.81	1.81	1.81	1.96	1.92	1.83
High Apt, high Per									
Males	1.73	2.00	1.87	2.20	2.00	2.00	2.27	2.33	2.05
Females	1.80	2.20	1.80	2.27	2.40	2.27	2.60	2.40	2.22
M	1.77	2.10	1.83	1.77	2.20	2.13	2.43	2.37	2.13

Note.—Matrices are listed in the order of their increasing effect of choice on the team outcomes. For low Apt groups, $N = 13$ males, 13 females; for high Apt groups, $N = 15$ males, 15 females.

TABLE 3

ANALYSIS OF VARIANCE SUMMARIES OF DATA INVOLVING LOW-D_{TI} MATRICES AND DATA INVOLVING HIGH D_{TI} MATRICES

Source	Low-D_{TI} matrices (2, 3, 4, 6)			High-D_{TI} matrices (1, 5, 7, 8)		
	df	*MS*	*F*	*df*	*MS*	*F*
Sex (A)	1	4.93	3.14	1	4.72	2.66
Aptitude (B)	1	1.91	1.22	1	2.08	1.17
Performance (C)	1	9.43	6.02*	1	5.14	2.90
A × B	1	.16		1	.79	
A × C	1	1.39		1	2.28	1.29
B × C	1	.60		1	8.23	4.64*
A × B × C	1	.00		1	3.28	1.85
Error (b)	104	1.57		104	1.77	
Matrices (D)	3	3.75	8.18**	3	2.28	5.68**
A × D	3	.34		3	.19	
B × D	3	.37		3	.14	
C × D	3	.10		3	1.55	3.85*
A × B × D	3	.54	1.17	3	.54	1.34
A × C × D	3	.40		3	.58	1.43
B × C × D	3	.16		3	.16	
A × B × C × D	3	.03		3	.32	
Error (w)	312	.458		312	.402	

Note.—*F*-ratios ≤ 1.0 are not shown.
* $p < .05$.
** $p < .01$.

lowest (Matrices 2, 3, 4, and 6) and (*b*) only the four matrices for which D_{TI} was greatest (Matrices 1, 5, 7, and 8). These analyses are summarized in Table 3.

Analyses of responses to low-D_{TI} matrices yielded significant effects of matrix type and performance but no significant interactions. Data relevant to these analyses, shown in Table 4a, indicate that high performers were significantly more team-oriented ($M = 1.89$) than were lower performers. This relationship was equally strong at both levels of aptitude.

Analyses of responses to high-D_{TI} matrices yielded a significant main effect of matrix type. However, unlike the analyses of low-D_{TI} data, performance interacted significantly both with aptitude and with type of matrix. Data relevant to these interactions are shown in Table 4b.

The significant interaction of aptitude and performance indicates that a positive relationship between performance and the number of team-oriented responses occurred only among students of high aptitude ($F = 7.38$, $p < .01$). In fact, high performers of high ability were more team oriented ($M = 2.28$) than were *S*s at any other level of aptitude and performance, and differed from *S*s in the other three cells combined ($M = 1.87$; $F = 8.15$; $p < .01$).

The significant interaction of performance and matrix type appears due to the fact that performance was related positively to the number of team-oriented responses to all matrices except Matrix 5, where the relationship was nonsignificantly negative. This matrix is similar to the

TABLE 4

MEAN NUMBER OF TEAM-ORIENTED RESPONSES ON MATRICES LOW IN D_{TI} AND MATRICES HIGH IN D_{TI} AS A FUNCTION OF APTITUDE, PERFORMANCE, AND SEX

Matrices	Males			Females		
	Low performance	High performance	*M*	Low performance	High performance	*M*
Low in D_{TI}						
High aptitude	1.48	1.95	1.72	1.77	2.02	1.89
Low aptitude	1.38	1.71	1.55	1.75	1.84	1.80
M	1.44	1.84	1.64	1.76	1.94	1.85
High in D_{TI}						
High aptitude	1.67	2.15	1.91	1.97	2.42	2.19
Low aptitude	2.06	1.65	1.86	1.84	2.10	1.97
M	1.85	1.92	1.88	1.91	2.27	2.09

well-known "prisoner's dilemma" in which, if both players try to maximize personal gain, they decrease both individual and group payoffs. Such a matrix may introduce additional factors into the decision making that distinguishes it from the others used.

It was speculated that differences between males and females in the relationship between performance and the frequency of team-oriented responses might occur as a result of sex differences in the relationship of academic achievement to social goal attainment. This hypothesis was not supported on the basis of the above data. When choices had little effect upon team outcomes, the frequency of team-oriented responses increased with performance among all *S*s. When the effect of choices upon team outcomes was greater, the number of team-oriented responses increased with performance level among high-aptitude students of both sexes but was not substantially related to performance among either males or females of low measured ability. Among males, the relationship was nonsignificantly negative.

Discussion

To summarize the major results of this study: When choices had relatively little effect upon team outcomes, the tendency to make responses that would increase the likelihood of team goal attainment was related positively to academic performance among both males and females. When decisions had a relatively great effect upon team outcomes, the tendency to make team-oriented choices was related positively to performance among high-aptitude students of both sexes, but was unrelated to performance among low-aptitude students.

The results of this study therefore call into question the general validity of the assumption that the likelihood of academic success is greater among students who are not socially motivated and who therefore are more apt to pursue academic goals without being distracted by competing social interests. While high academic performance may often require concentrated independent effort, the tendency to cooperate with others in mutual pursuit of achievement goals may generalize to the academic environment and may result in an increase in academic effectiveness. To this extent, social orientation may actually facilitate achievement in college.

The only qualification to the interpretation offered above concerns overachieving males (high performers of low measured ability). These *S*s were less team-oriented than low performers of low ability when the effect of their choices on team outcomes was high. Furthermore, supplementary analyses of individual correlations between team orientation and D_{TI} indicated that male overachievers, unlike *S*s in any other academic category considered in this study, actually decreased their team orientation slightly when team outcomes were more affected. It may be speculated that overachievement among males results in part from a general desire to be recognized as an individual for success in goal-directed activity. Overachievement among females, however, would not have similar roots.

Students who perform poorly despite high ability might be expected to show little interest in being recognized for personal achievement. They nevertheless appear to prefer to seek goals independently of other persons. In this regard, Wyer (1967) reported that when an incentive to perform well on a judgmental task was provided, underachievers of both sexes conformed less to group estimates than did students at any other level of aptitude and performance. This finding supports the view that underachieving students typically prefer not to rely upon others in achievement-related activity.

Other interpretations of the results of this study are plausible. In this regard, although high-ability, high-performing students of both sexes were relatively team-oriented, a possible distinction between males and females may be worth considering. Wyer (1967) found that when the incentive to perform well on a judgmental task was minimized and social group attractiveness increased, high-aptitude,

high-performing males conformed more to fictitious group norms than did males at other levels of Apt and Per, but high-aptitude, high-performing females conformed less than did females not fitting this description. Conformity under such conditions may be due primarily to concern over being accepted by other group members (Deutsch & Gerard, 1955; Wyer, 1966). Therefore, while the team orientation of high-performing males of high ability may be due in part to their concern over being liked or accepted by their partners, similar behavior among high-performing females of high ability may have different determinants. Females may have generally less interest in personal achievement than males. Those who perform well academically may be satisfied with the recognition they have received for individual achievement, and therefore may tend less to seek recognition in nonacademic situations.

Some of the fundamental questions concerning the motivational and behavioral correlates of academic achievement are still unanswered by the present study. For example, the behavioral characteristics of students in achievement-related activity that does not involve other persons are yet to be delineated. The validity of assumptions that underachievers have less interest in recognition for personal achievement than other persons, or that overachievers are more apt to persevere in achievement-related activity, has not been tested directly. More specific questions were also raised by this study which should be considered in further research in this area. Specifically, differences between matrices in the relative effect of choices upon team and individual goals were due primarily to differences in the effect of choices upon team outcomes. Since relationships involving performance were contingent upon the type of matrix involved, the use of matrices in which choices affect individual outcomes to a relatively greater extent than team outcomes should be considered. It may also be fruitful to explore choice behavior under conditions in which the likelihood that partners would be expected to make team-oriented choices is systematically varied. Finally, all groups used in this study were homogeneous with respect to sex; situations in which partners are of the opposite sex might produce substantially different results from those reported here.

REFERENCES

DEUTSCH, M. Cooperation and trust: Some theoretical notes. In M. R. Jones (Ed.), *Nebraska symposium on motivation: 1962.* Lincoln: University of Nebraska Press, 1962. Pp. 275–320.

DEUTSCH, M., & GERARD, H. B. A study of normative and informational social influences upon individual judgment. *Journal of Abnormal and Social Psychology,* 1955, **51,** 629–636.

LAVIN, D. E. *The prediction of academic performance,* New York: Russell Sage Foundation, 1965.

PIERCE, J. V., & BOWNMAN, P. H. Motivation patterns of superior high school students. In M. Kornrich (Ed.), *Underachievement.* Springfield, Ill.: Charles C Thomas, 1965. Pp. 214–252.

TODD, F. J., TERRELL, G., & FRANK, C. E. Differences between normal and underachievers of superior ability. *Journal of Applied Psychology,* 1962, **46,** 183–190.

WYER, R. S. Self-acceptance, discrepancy between parents' perceptions of their children, and goal-seeking effectiveness. *Journal of Personality and Social Psychology,* 1965, **2,** 311–316.

WYER, R. S. Effects of incentive to perform well, group attraction and group acceptance on conformity in a judgmental task. *Journal of Personality and Social Psychology,* 1966, **4,** 21–26.

WYER, R. S. Behavioral correlates of academic achievement: conformity under achievement and attraction incentive conditions. *Journal of Personality and Social Psychology,* 1967, **6,** 255–263.

WYER, R. S., & TERRELL, G. Social role and academic achievement. *Journal of Personality and Social Psychology,* 1965, **2,** 117–121.

WYER, R. S., WEATHERLEY, D., & TERRELL, G. Social role, aggression and academic achievement. *Journal of Personality and Social Psychology,* 1965, **1,** 645–649.

(Received March 8, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 2, 82-87

COGNITIVE CONFLICT AND REVERSIBILITY TRAINING IN THE ACQUISITION OF LENGTH CONSERVATION

FRANK B. MURRAY
University of Minnesota

A child conserved the length of 2 equal sticks that were made to look unequal by the Müller-Lyer illusion if he saw the illusion and despite it, maintained that the sticks were the same length. Nonconservers who were trained by a reversibility and cognitive conflict procedure did significantly better than untrained nonconservers on the Müller-Lyer task and on a transfer task with the same sticks distorted by the Oppel inverted T illusion ($p < .01$). There was no significant difference between the trained and untrained nonconservers on a transfer task in which equal areas were distorted by the Jastrow illusion.

Whatever the origins of our conception of length we are deeply convinced, as Einstein (1957) observed, that the length of a thing is constant even if its position is changed. The notion of the conservation of length refers to just this deep conviction, namely the recognition that the length of a thing is variant only under the transformations or operations of addition and subtraction, or in some cases from expansion or contraction due to temperature change or the sheer mechanical forces of stretching and compression. Length is invariant or conserved under such irrelevant transformations as position change, movement, filled space, perceptual distortion, etc.

Piaget, Inhelder, and Szeminska (1960) found, in studying the conservation of length, that when two identical sticks were evenly placed one under the other, children correctly judged that the sticks were the same length. However, when one of the sticks was moved slightly to the right of the other, children under the age of 8 generally thought the moved stick was longer than the unmoved one, that is they failed to conserve length in this instance. Children under the age of 8 also judged a straight line and sinuous one to be equal in length when the lines' end points were even. These children, moreover, persisted in claiming that the lines were equal after the sinuous line had been stretched to show its greater length and returned to the original position with its end points flush with those of the straight line. Children also judged two unequal lines, straight and curved, whose end points were all even, to be the same length. That length was not conserved by children under 8 in these instances was found in a replication by Lovell, Healey, and Rowland (1962).

It might have been the case that in the above studies the children and the experimenter meant different things by length, the former denoting by it only the position of the lines' end points, a strategy that would, more often than not, produce successful encounters with the different lengths of things in the child's world. A previous study (Murray, 1965) has shown that if the equality of two lengths whose end points are even is distorted by some common geometrical illusions, nonconservation is still found in children under 8, and therefore must result from more than just the misconception of length based only on the relative position of the extremities of lines.

The phenomenon of nonconservation of length itself is surprising in children nearly 8 years old, particularly since average children between 3 and 3½ years of age can successfully (82% of the time) follow verbal instructions to pick out the longer of two sticks on the Binet test. More surprising than the phenomenon of nonconservation itself, however, is the finding in conservation learning or training studies, that even the most sound and reasonable training procedures "have had remarkably little success in producing cognitive change [Flavell, 1963, p. 378]." The evidence from training studies of the conservation of

number (Wohlwill & Lowe, 1962), conservation of substance and weight (Smedslund, 1961a, 1961b, 1961c), conservation of area (Beilin & Franklin, 1962) and conservation of length (Smedslund 1963a, 1963b) indicates that a child cannot be taught the concept in question unless he has developed the particular cognitive structure in question and this structure does not seem significantly altered by external reinforcement or manipulation.

Smedslund (1961d) proposed that a state of cognitive conflict was the precursor of the cognitive reorganization that was required to support conservation. The proposal is consonant with the Piagetian notion that problems provoke cognitive disequilibrium, the resolution of which requires a new integration of distinct operations, such as simultaneously, instead of separately, attending to height and width in substance conservation or both end points in length conservation. Smedslund (1961d) created the conflict by changing the shape of a clay ball to make it appear larger when a piece had been taken away to make it actually lighter. Smedslund (1963b) has had mild success with this method of the juxtaposition of competing forces in training the conservation of length. Gruen (1965) has used the technique in training number conservation and found it was successful only when it was coupled with verbal pretraining. The cognitive conflict technique was used in the present study to the extent that subjects (*S*s) were directed to perform actions that made the same stick appear longer and sometimes shorter than an equal length stick.

The technique was based primarily on the notion that the cognitive operations that support conservation are reversible, for example, the operation that made the equal length sticks unequal can be undone by the inverse operation. To conserve any property the child presumably needs the rule that allows him to get from the original state to the transformed state and back again. The notion of reversibility is for Piaget the defining property of the cognitive operations that support the conservations. Wallach and Sprott (1964) have been successful in training number conservation with a reversibility procedure, but Beilin (1965) was unsuccessful in training number and length conservation with it. The reversibility procedure used in this study produced a conflict between the original and transformed states that could be resolved by allowing *S* to perform the actions that connected the two states.

METHOD

Subjects

All 119 *S*s, 69 boys and 50 girls, were enrolled in the kindergarten, first, and second grades of the Lida Lee Tall Laboratory School in Towson, Maryland. The mean age for each grade was 5.76 years (SD = .30, range 5.25–6.16), 6.71 years (SD = .33, range 6.25–7.16), and 7.79 years (SD = .34, range 7.25–8.16) respectively. The mean IQ from the SRA Primary Mental Abilities Test was 112.4 (SD = 11.4) for the first grade and 112.5 (SD = 10.9) for the second grade. Test results for kindergarten were not available. There were in all 31 kindergartners, 42 first, and 46 second graders.

All testing was done individually in a small conference room in the laboratory school. Each *S* faced the experimenter across a table, which held the stimulus materials used in the study.

Materials

The materials for the length conservation test were: (*a*) 10 round white sticks, 5 millimeters in diameter, two of which were the same length (21 centimeters), four were unequally longer than 21 centimeters and the rest, shorter, (*b*) a black rectangular (30 × 50 centimeters) composition board on which were glued three cardboard figures—the arrows and feathers of the Müller-Lyer figure in horizontal position (see Figure 1). The two segments of the arrows and feathers were 5 millimeters × 7 centimeters and at right angles to one another.

Procedure

The experimenter introduced all tasks as a game, and explained his notetaking as keeping score so he wouldn't forget how well *S*s did. Before the test of length conservation, *S*s were presented a long and a short stick from the 10 sticks described above, and asked to find the longer one. The *S*s were asked then if they could find two sticks in the 10 that were the same length or size. If *S* had difficulty with the multiple discriminations, the experimenter prompted by pointing to two unequal sticks and asking about their length, and then pointing to the two equal sticks and asking about their length. The children's responses to these preliminary questions ensured that *S*s understood the directions in the way that the experimenter intended.

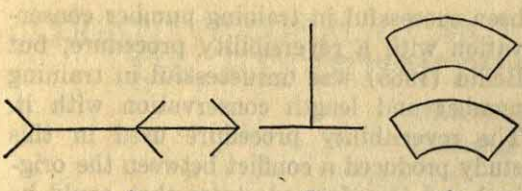


FIG. 1. The Müller-Lyer illusion (A), Oppel inverted T illusion (B), and the Jastrow illusion figures (C).

The experimenter took the equal sticks that *S* had selected and placed them upright side by side on the table to enable *S* to reconfirm their equality, before the experimenter slowly placed them in the Müller-Lyer configuration to make their lengths appear unequal. One-half *Ss* in each grade had the "feathered figure" on the left, and the rest had it on the right. The *Ss* were asked:

> Does this stick (experimenter pointed to the stick in the "feathered" configuration) *look* longer than this stick (experimenter pointed to the "arrowed" stick), does it *look* the same length as this stick, or does it *look* shorter than this stick?

To nullify any leading influence the order of the alternatives (longer, shorter, the same length as) might have on *S*'s response, the order of the alternatives was varied randomly among *S*. If *S* answered that the sticks looked the same length, that is, he failed to see the illusion, he was discarded from the experiment. The *Ss* were asked a second question:

> If the sticks were standing up the way they were before, would this stick (experimenter pointed to the stick in the feathered configuration) be longer than this stick (experimenter pointed to the other stick), would it be the same length as this stick, or would it be shorter than this stick?

Again, the order of the alternatives was varied randomly among *Ss*.

One week from the session in which the above procedures and others described in Murray (1967a, 1967b) were administered, *Ss* who were scored as nonconservers on the Müller-Lyer problem were again given that length conservation test. All *Ss* who could still be scored as nonconservers were randomly divided within each grade into a control group and an experimental group that would receive training in length conservation.

The control group *Ss* met individually with the experimenter and were given the following tests: (*a*) the length conservation on the Müller-Lyer figure described above, (*b*) length conservation test in which the sticks were held by the experimenter in the form of the Oppel inverted T illusion, and (*c*) a size or area conservation problem in which two equal ring segments were held by the experimenter in the pattern of the Jastrow illusion (Figure 1).

In the inverted T problem, *Ss* were asked (*a*) if the vertical stick looked longer, etc., than the horizontal one (experimenter pointed to both sticks), and (*b*) if it was really longer, etc. In the Jastrow figure, *Ss* were asked similar questions about the apparent and real size differences between the bottom and top ring segments.

Before the experimental group was subjected to the three conservation tests given to the control group, they were trained in the following manner: (*a*) the sticks in the Müller-Lyer figure were placed by the experimenter before *Ss* as before, but *S* was allowed to pick them up to confirm their equality and himself replace them in the distorting figure, (*b*) *S* was permitted to pick up and replace the sticks two more times, (*c*) *S* was directed to try switching the sticks, that is, replacing the one from the feathered configuration in the place of the one from the arrowed configuration and vice versa, and (*d*) *S* was permitted to switch the sticks two more times.

RESULTS

All *Ss* were able to answer the preliminary questions about unequal and equal length correctly. Each *S* was scored as having conserved length only if (*a*) he saw the Müller-Lyer illusion, that is, said the feathered stick looked longer, and (*b*) despite seeing the illusion maintained that the sticks were the same length. If *S* maintained that the sticks were unequal in length, he was scored, of course, as a nonconserver.

A median test (Seigel, 1956) was applied to the scores, and indicated a significant difference in conservation for the group above the median age (6.91 years) and for the group below it ($\chi^2 = 14.4, p < .001$).

Chi square for differences in conservation between males and females was insignificant for the group ($\chi^2 = 1.98, p < .20$). Thirty-seven of the 38 *Ss* identified as nonconservers on the length conserva-

TABLE 1

NUMBER OF CONSERVERS AND NONCONSERVERS ABOVE AND BELOW THE GROUP MEDIAN AGE OF 6.91 YEARS

Group	Conservers	Nonconservers
Above	51	9
Below	30	29

Note.—Range 5.25–8.16 years.

TABLE 2

NUMBER OF CONSERVERS (C) AND NONCONSERVERS (NC) IN LENGTH CONSERVATION AND TRANSFER OF LENGTH CONSERVATION TASKS BETWEEN TRAINING AND CONTROL GROUPS

Groups	Müller-Lyer task*		Transfer tasks			
			Oppel T*		Jastrow**	
	C	NC	C	NC	C	NC
Experimental	14	1	11	4	7	8
Control	0	14	1	13	0	14

Note.—For the experimental group, $N = 15$; for the control group, $N = 14$.
* $p < .01$.
** $p > .10$.

tion task were retested on the same task. Twenty-nine were still found to be nonconservers and eight were found to be conservers. The 29 *S*s randomly divided by grade constituted the control and training groups.

On the Müller-Lyer length conservation task (Table 2) there were by the Fisher Test (Siegel, 1956) a significantly greater number of conservers from the training group than the control group ($p < .01$). On the test of length conservation transfer to the Oppel inverted T task, there were also by the Fisher Test a significantly greater number of conservers in the training group than the control group ($p < .01$), but on the Jastrow size or area conservation test the difference in conservers between the training and control groups was insignificant ($p > .10$). The significance levels are two-tailed and conservative since a greater number of conservers was predicted in the group that had training.

DISCUSSION

The transition from nonconservation to conservation was found to be between 6 and 7 years of age. Comparisons of this finding with age norm findings in other studies must consider at least the following variables which have been found to be related to conservation: IQ of *S*s, task complexity and materials, field-independence of *S*s, and socioeconomic status, culture, and urban-rural location. In a previous study (Murray, 1965) the transition from nonconservation to conservation occurred between 7 and 8 years. The most parsimonious explanation for the higher transition age found in that study is that older children were *S*s. Arguments could be advanced, however, that the length conservation task used in the present study was simpler and more concrete than the tasks used in the previous study.

Piaget, Inhelder, and Szeminska (1960) found that length conservation was complete for 50% of those children aged 7–7½ years, and Vinh-Bang's data (Smedslund, 1963a) sets the 50% conservation level closer to 8 years of age. In the present study, 52% of the 5- and 6-year olds (median age = 6.16), 72% of the 6- and 7-year old (median age = 6.91) and 86% of the 7- and 8-year olds (median age = 7.66 years) conserved length. It can be said that Piaget's findings on the relationship between conservation and age were broadly confirmed.

Although precautions were taken in the present study to ensure that *S*s knew the sticks were the same length, it might be argued that this knowledge was forgotten. Lovell and Ogilvie (1960) point out, however, that forgetting is not a critical factor in nonconservation, since many conservers could remember the original situation.

It might be argued also that conservation or nonconservation was found in the present study simply because the second question, namely, "How is it really?" suggested to *S*s that a different answer was being looked for. McConnell (1963) has found, for example, that when children judge the relative sizes of two equal geometrical forms, they judge the forms to be unequal simply because the opportunity suggested itself in the question. McConnell found the suggestibility factor to be negatively related to age (6–18 years), although the curvilinear relationship showed no appreciable differences between 6-, 7-, and 8-year olds. If such a factor were operating in the present study, it would tend to lower the nonconservation-conservation transition age, by lowering the number of

nonconservers, namely, *S*s who gave the same answer to both questions. If it were the sole factor operating on the responses to the second question, one-half of the *S*s at each age would be scored conservers; a binomial test failed to support the hypothesis.

It is possible that by answering the second question indiscriminately, one-third of the *S*s could have conserved the item by chance alone. A binomial test ($p = 1/3$) on the proportions of nonconservers and conservers at each age indicated that only the youngest *S*s (below age 6.16 years) could have conserved length in this problem by guessing alone.

As a result of reversibility and cognitive conflict training, nonconservers acquired conservation in the Müller-Lyer task they were trained on, and in the Oppel inverted T task on which they were not trained. Training did not seem to be effective in producing transfer to the area or size conservation task. The lack of transfer in trained conservation was found also by Gruen (1965) in which training in number conservation did not transfer to length and substance conservation, and by Beilin (1965) who found that training in length and number conservation did not transfer to area conservation. It would seem that the conservations are acquired for each concept separately, although it could be, as Smedslund (1961b) has shown, that conservation that is laboratory acquired may not have as much depth or generality as conservation that is "naturally" acquired.

Beilin and Franklin (1962) and Beilin (1965) have found that training in conservation was more effective with older nonconservers than the younger ones. The success of the present training procedure may have resulted in part from the use of older *S*s, though the median ages of the *S*s in the present study and in Smedslund's study (1963b) were virtually the same. Nevertheless, the age of nonconservers (median age = 6.16 years) in the present study was close to the age that length conservation would begin to occur "naturally."

Other reasons for the effectiveness of the training procedure can be speculated on. Since forgetting the equality is not a significant factor in nonconservation, the child may lack an awareness of the possibility of return to the original situation. The child considers the two states to be static and separate, because the rules or action relating them are absent. Inhelder (1965) cites evidence to show that young preoperational children were unaware, for example, of the successive forms an arc would take in being straightened, while those that conserved could represent the intermediate stages of the transformation from arc to straight line. It could be argued that the present training procedure, by allowing *S* to actively manipulate the sticks from one state to the other, facilitated *S*'s awareness of the relationship between the equal and unequal appearing states.

The phenomena of nonconservation and conservation themselves are important facts in the psychology of subject matter. The fact that instruction may only be effective when children are close to the time at which length is conserved naturally has obvious implications for the educational notions of readiness and curriculum sequence. The conserved concept of length, important in itself, is a requisite for the concepts of transitivity and measurement. It is clear from the present study that nonconservers can be taught to conserve; nevertheless, premature instruction in the concept may result in no more than the acquisition of verbal fluency which masks a conceptual defect that has the magnitude of nonconservation itself. Additional instructional techniques, not unlike the one used in the present study, certainly can be devised to influence cognitive development and make its sources more explicit.

REFERENCES

BEILIN, H. Learning and operational convergence in logical thought development. *Journal of Experimental Child Development*, 1965, **2**, 317-339.

BEILIN, H., & FRANKLIN, I. Logical operations in area and length measurement: Age and training effects. *Child Development*, 1962, **33**, 607-618.

EINSTEIN, A. *Relativity, the special and the general theory*. (15th ed.) London: Methuen, 1957.

FLAVELL, J. *The developmental psychology of Jean Piaget*. New York: Van Nostrand, 1963.

GRUEN, G. Experiences affecting the development of number conservation in children. *Child Development*, 1965, **36**, 963–980.

INHELDER, B. Operational thought and symbolic imagery. *Monographs of the Society for Research in Child Development*, 1965, **30**, 4–18.

LOVELL, K., HEALEY, D., & ROWLAND, A. Growth of some geometrical concepts. *Child Development*, 1962, **33**, 751–767.

LOVELL, K., & OGILVIE, E. A study of the conservation of substance in the junior school child. *British Journal of Educational Psychology*, 1960, **30**, 109–118.

McCONNELL, T. Suggestibility in children as a function of chronological age. *Journal of Abnormal and Social Psychology*, 1963, **67**, 286–289.

MURRAY, F. B. Conservation of illusion-distorted lengths and areas by primary school children. *Journal of Educational Psychology*, 1965, **56**, 62–66.

MURRAY, F. B. Conservation of illusion-distorted length and illusion strength. *Psychonomic Science*, 1967, **7**, 65–66. (a)

MURRAY, F. B. Some factors related to the conservation of illusion-distorted length by primary school children. *1967 AERA Proceedings*, 1967, 193–194. (b)

PIAGET, J., INHELDER, B., & SZEMINSKA, A. *The child's conception of geometry*. New York: Basic Books, 1960.

SIEGEL, S. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill, 1956.

SMEDSLUND, J. The acquisition of conservation of substance and weight in children. II External reinforcement of conservation of weight and of the operations of addition and subtraction. *Scandinavian Journal of Psychology*, 1961, **2**, 71–84. (a)

SMEDSLUND, J. The acquisition of conservation of substance and weight in children. III Extinction of conservation of weight acquired "normally" and by means of empirical controls on a balance. *Scandinavian Journal of Psychology*, 1961, **2**, 85–87. (b)

SMEDSLUND, J. The acquisition of conservation of substance and weight in children. IV Attempt at extinction of the visual components of the weight concept. *Scandinavian Journal of Psychology*, 1961, **2**, 153–155. (c)

SMEDSLUND, J. The acquisition of conservation of substance and weight in children. V Practice in conflict situation without external reinforcement. *Scandinavian Journal of Psychology*, 1961, **2**, 156–160. (d)

SMEDSLUND, J. Development of concrete transitivity of length in children. *Child Development*, 1963, **34**, 389–405. (a)

SMEDSLUND, J. Patterns of experience and the acquisition of conservation of length in children. *Scandinavian Journal of Psychology*, 1963, **4**, 257–264. (b)

WALLACH, L., & SPROTT, R. Inducing number conservation in children. *Child Development*, 1964, **35**, 1057–1072.

WOHLWILL, J., & LOWE, R. An experimental analysis of the development of the conservation of number. *Child Development*, 1962, **33**, 153–167.

(Received March 13, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 2, 88-93

"OVERPROMPTING" IN PROGRAMMED INSTRUCTION[1]

RICHARD C. ANDERSON, GERALD W. FAUST, AND MARIANNE C. RODERICK
University of Illinois

108 Ss completed either the Standard version or a Heavily-Prompted version of a 1,052-frame section of a psychology program. Half of the Ss made constructed responses while the remainder were instructed to "think" the answers that went into the blanks. As expected, those who received the Standard program scored higher on the posttest and took longer to complete the program than those who received the Heavily-Prompted version. Response mode made no difference. The results were interpreted as showing that arrangements of lesson material which permit the student to respond correctly without noticing the cue undermine performance.

One of the stocks in trade of programmers is prompting, a technique of providing information that helps the student to give a correct answer. Depending upon the context, a prompt may consist of a rule which can be applied to an example, a hint to help in the solution of a problem, the first letter of an answer, a synonym for an answer, or one of many other devices both simple and complex. A *prompt* is technically defined as a stimulus that already controls or partially controls a response. The instructional problem is to arrange a shift in stimulus control from the prompt to the discriminative stimulus or *cue*, a second stimulus (or set of stimuli) which prior to instruction does not control the response. The shift in stimulus control has been accomplished when the student can make the response when the cue alone is present.

Much of the actual research on prompting has entailed paired-associate lists. It has been repeatedly demonstrated that people learn faster under a prompting procedure, in which both the stimulus term and response term appear before the response is required, than under the anticipation method, or confirmation method as it has been called in these studies (e.g., Cook & Spitzer, 1960; Levine, 1965; Sidowski, Kopstein, & Shillestad, 1961). On the basis of these experiments, Cook (1963) has been willing to extrapolate to instructional practice. He has argued that the correct answer should be indicated to the student *before* he makes his response. The implication is that there is no such condition as "overprompting," a state of affairs, warned against in programmed instruction manuals, in which the response is said to be too completely determined by the prompt.

We contend that under certain conditions stereotyped and repetitious use of prompted frames can impair the effectiveness of programmed materials. The argument is that many students will begin to respond on the basis of the prompt alone, when the design of the frame permits it. As a result, their behavior will not come under the control of the cues because they will not have paid attention to the cues. The deleterious effects of an "overprompted" program would be expected to be most pronounced when, for instance, the students are bored, tired, or eager to finish quickly.

Anderson and Faust (1967) developed a stylized Russian vocabulary program in which each frame consisted of a paragraph of five sentences with English subjects and Russian predicate nominatives. Immediately below the paragraph one of the sentences was repeated with a blank in place of the Russian word. In a second otherwise

[1] The authors are indebted to Thomas Anderson for assistance in the development of the posttest; to Jean Behrens for assistance in data collection; to Richard Spencer and Elwood Leslie for help in machine scoring answer sheets and completing item analyses; to the McGraw-Hill Book Company for permission to reproduce the program; and to James Holland for providing materials to teach people how to use the blackout technique. A version of this paper was read at the American Educational Research Association annual meeting, New York, 1967.

identical version of the program the Russian word which was to serve as the response was underlined in every frame. Each frame in the No-Underline version required the student at least to notice the cue in order to respond correctly, since the only way to locate the Russian word that went in the blank was to find the English word with which it was associated. The Underline version, on the other hand, permitted the student to copy Russian words into the blanks without ever looking at the English words. As predicted, although both versions led to error-free performance during the program, on the posttest students who received the No-Underline version recalled significantly more Russian words practiced during the program than did students who received the Underline version.

The chief purpose of the investigation described in this paper was to determine whether the analysis of overprompting which has just been described applies to actual lesson materials. At least one study suggests that it does. Kress and Gropper (1966) presented a 43-frame program on "elasticity" and a 44-frame program on "direct and inverse relationships" to eighth graders over closed-circuit television. Some groups saw the Nonprompted versions. Other groups saw the Partially-Prompted programs, in which the first letter of every response word was supplied. The remainder received the Fully-Prompted versions, in which the responses were printed entirely in capitals. The results on the posttests showed that the more prompting the lower the achievement.

Of course, in practice, prompts are seldom used in quite such a heavy-handed and routinized fashion as they have been in the experiments on prompting that have been completed thus far. It remains to be seen whether similar results can be obtained when a variety of strong prompts are employed in a relatively unobtrusive manner.

A secondary objective of the present study was to investigate the joint effects of amount of prompting and response mode.

METHOD

Subjects

One hundred and eight students, mostly secondary school teachers, enrolled in a summer course in educational psychology served as subjects (*Ss*). Participation was a course requirement for 63 *Ss* whereas 45 were volunteers who had the option of participating in the experiment in lieu of writing a paper. The *Ss* were randomly assigned to groups with the constraint that each contain proportional numbers of volunteers and nonvolunteers. A total of 12 other *Ss* were discarded for the following reasons: eight because they indicated (on a questionnaire completed after the fifth program set) previous exposure to the program or considerable knowledge of the material contained in the program; two because they failed to complete the experiment; and two in order to equalize numbers in the treatment groups.

Materials

Two versions of the first 1,052 frames (25 sets) of the Holland and Skinner (1961) program *The Analysis of Behavior* were prepared. The Standard version was identical to the original, except for the physical arrangement of frames, while the Heavily-Prompted version was altered to include one additional prompt on about 90% of the frames. The remainder of the frames were already heavily prompted. Prompts were introduced according to the following rules: (*a*) the response term was always underlined in copying frames; (*b*) the appropriate article was used before each response blank and ambiguity about whether the response was plural or singular was removed; (*c*) in multiple-choice frames the first alternative was always correct; (*d*) strong connectives (e.g., therefore) were added when these emphasized existing prompts; and (*e*) when nothing else seemed possible the first letter or two of the response was provided.

The programs were mimeographed on 16-pound white stock and placed in booklets, one booklet for each set, with 3½ × 8½-inch pages stapled along the left margin. Each page contained a centered frame above which was the feedback for the preceding frame. The exhibits were in a separate booklet composed of longer pages, one exhibit per page.

Procedure

The experiment was conducted in a large, air-conditioned room containing space for 15–20 people at long tables. Free coffee was provided. Each *S* scheduled himself to work on the program in 1-hour blocks over a 1-month period, determining for himself how his time would be distributed. A library arrangement was employed wherein *S* checked out a single booklet at a time and returned it when completed to a "librarian," who checked

for proper identification and then issued the next booklet if *S* wanted to continue.

The *S*s paced themselves; however, they did record the time on a cover sheet immediately before beginning a set and again on an end sheet immediately after completing the set. Since no one was allowed to stop in the middle of a set, training time was measured rather accurately.

A questionnaire was administered immediately after *S* completed the last booklet to assess his attitude toward programmed instruction and to determine the way in which he went about completing the program. Finally, a two-part posttest was given approximately 24 hours after *S* completed his last booklet. The posttest consisted of 20 short-answer questions, which were given first, and 51 multiple-choice items. The total posttest score consisted of the score on the short-answer section plus the score on the multiple-choice section corrected for guessing. Prior to the experiment the test was revised on the basis of data from several groups of *S*s, both naïve and sophisticated. Special attention was paid to the development of discriminating items for high achievers.

Results

As can be seen from Table 1, which contains analyses of variance, and Table 2, which contains means and standard deviations, the main hypothesis was confirmed. Those who received the Standard program learned significantly more than those who got the Heavily-Prompted version.

Considering the fact that a relatively large number of *S*s participated, the fact that a lengthy and, it was imagined, discriminating achievement test was used, and especially the fact that a rather long program was employed, it was fully expected that for those who received the Standard program the experiment would show a significant posttest advantage for overt responding. The fact that it did not ($t = 1.07$) gives us pause. Two other studies (Holland, 1965; Williams, 1963) which employed sections of the same program have found an advantage for overt responding. There were several differences in procedure between the earlier experiments and the present one that may account for the difference in results. First, in the Holland experiment and the Williams experiment *S*s read the frames with the blanks filled in whereas in this experiment *S*s in the covert condition were instructed to "think" the answers that went in the blanks. Perhaps the latter procedure is somewhat more effective than reading alone. Second, all of the *S*s in the preceding experiments were required to participate in order to fulfill course requirements while about 40% of the *S*s in the present experiment were volunteers. Among those who took the Standard program, volunteers who made overt responses achieved a posttest mean of 24.2 and volunteers who were instructed to make covert responses achieved a posttest mean of 24.1. In contrast, among those who took the Standard program and were required to participate, *S*s who made overt responses showed a total posttest mean of 26.4 and those who made covert responses showed a mean of 20.1. While the latter comparison was still not significant ($t = 1.34$), it did seem to make a difference whether or not *S* was a volunteer. There is no reason to believe that there is a value to overtness for its own sake. Covert responses are fine as long as they occur. The problem is that covert responding may drop out after a period of time. Apparently volunteers sustained covert responding to a greater extent than nonvolunteers.

Previous research (Williams, 1963, 1965) has indicated that an advantage from requiring overt, constructed responses will be found primarily when the posttest requires *S* to supply novel, technical terms. In the present study there was no difference between the overt and covert groups on the short-answer section of the posttest, every question of which required *S* to supply a technical term ($t = .00$). Nor was there a significant difference between the overt and

TABLE 1
Analysis of Variance of Training Time and Posttest Performance

Source	*df*	Training time		Posttest performance	
		MS	*F*	*MS*	*F*
Prompting (P)	1	248,256	36.54**	773.3	4.75*
Response mode (R)	1	449,823	66.22**	114.1	<1
P × R	1	14,701	2.16	73.3	<1
Residual	104	6,793		162.7	
Total	107				

* $p < .05$.
** $p < .01$.

TABLE 2

TRAINING TIME AND POSTTEST PERFORMANCE MEANS AND *SD*s

Group	Training time (in minutes)		Posttest performance	
	M	*SD*	*M*	*SD*
Standard overt	470	114	25.3	13.5
Standard covert	317	64	21.6	11.6
Heavily-prompted overt	350	79	18.3	12.1
Heavily-prompted covert	245	61	17.9	13.6

covert conditions on the short-answer test considering only *S*s who took the Standard version of the program ($t = .86$). Incidentally, there was a larger difference between the two program versions on the short-answer section of the test ($F = 10.69$, $df = 1/104$, $p < .01$) than on the multiple-choice section ($F = 2.60$, $df = 1/104$, $p > .05$).

None of the posttest effects was quite as strong as had been expected. To our distress we have discovered a possible reason for weak effects. On many frames it was possible to read the correct answer (which appeared on the top of the next page) through the paper. This was a prompt that we had not counted on, and it undoubtedly reduced the difference between the Standard and Heavily-Prompted versions (see Brown, 1966). This unplanned prompt may also have reduced the difference between the overt responding and covert responding conditions for those who received the Standard program. One of the items in the questionnaire asked *S* to describe any shortcuts he used in completing the program. Perhaps it is significant that five of the eight *S*s who reported that they occasionally read the correct answer through the paper were in the Standard Overt group. One *S* expressed the problem well. He wrote,

> I was annoyed at times by the fact that the correct responses on the next page were often visible because of the thiness (sic) of the paper. In cases where I had already looked at these responses it's difficult to tell if I really thought out & wrote down the correct response, or just copied the response that was visible through the page.

As one might expect, the error rate was lower in the Heavily-Prompted version (4.1%) than in the Standard version (8.0%) of the program ($t = 3.08$, $df = 52$, $p < .01$).

Figure 1 pictures the time in seconds per frame to complete the program sets for each of the treatment groups. Since the record of training time included time spent reading exhibits, four undergraduate volunteers were hired to read the exhibits for meaning in order to provide a basis for adjusting work rates. Each *S* was timed individually on each exhibit with a stop watch. The results, expressed in seconds per frame, appear in the bar graph at the bottom of Figure 1. The figure reveals that work rates were very stable for every treatment group across the entire program.

The thesis is that the Heavily-Prompted version resulted in poorer achievement than the Standard program because it permitted attenuated inspection behavior. It is argued that students often were able to respond correctly on the basis of prompts without paying attention to the entire cue, even, for example, to the definition of a technical term. As a result the cue sometimes failed to become a discriminative stimulus for the response and so, for instance, the student was sometimes unable to produce a technical term when its definition appeared on the posttest. The time data furnished indirect evidence that this analysis is correct, if it is assumed that the additional time spent by those who received the Standard program must have been reading time and thinking time. More direct evidence was obtained from the questionnaire. Three open-ended, essay-type questions asked the student to comment on the procedure employed to complete the program, including any shortcuts used. Not counting those who said they sometimes read the answer through the page, there were 22 *S*s who reported using prompts as shortcuts for filling blanks, distributed among treatment groups as follows: Standard Overt, 0; Standard Covert, 1; Heavily-Prompted

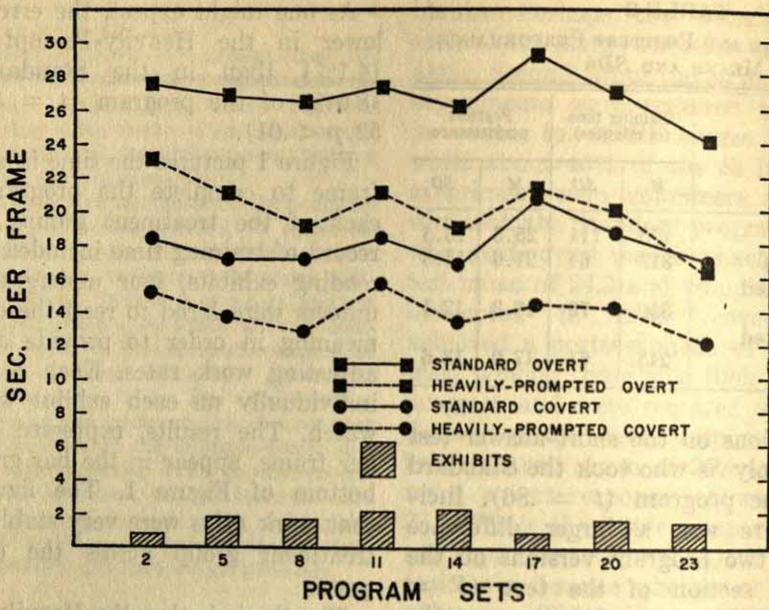


FIG. 1. Training time per frame. (Each point represents pooled data from adjacent sets.)

Overt, 11; Heavily-Prompted Covert, 10. The difference between the Standard and Heavily-Prompted versions was significant ($\chi^2 = 22.8$, $df = 1$, $p < .001$). The comments themselves were often informative. One *S* wrote, "When the first letters of the missing answer were given, it was difficult to not answer [before] studying the question first, as the answers were obvious." Another commented, "I noticed that the underlined words were usually the answers so [I] often copied them before reading." Still another explained, "By seeing the first letter or letters of the word I immediately wrote down the word without understanding fully the written material."

The present experiment complements the important work of Holland and Kemp (1965; Kemp & Holland, 1966) who have maintained that the effectiveness of a program depends upon the extent to which responses are contingent upon attention to the "critical material" (presumably meaning the set of stimuli that we have called the cue) in the frames. They have developed a procedure called the "blackout technique" to measure this contingency. The "blackout ratio" is the proportion of material in a program that can be lined through with a black crayon without affecting error rate. Two undergraduates were hired to apply the blackout technique to the programs used in this experiment. Beforehand they received about 20 minutes of training in the technique using materials prepared by Holland and Kemp for this purpose. Neither rater was familiar with the program or the subject matter and both were ignorant of the fact that there were actually two programs instead of one. The first four sets of the program were processed. In order to control individual differences in tendency to eliminate material, one rater evaluated the first and third sets from the Standard program and the second and fourth sets from the Heavily-Prompted version; the assignment was reversed for the other rater. The blacked out programs were not tried to determine if error rate had been affected. The blackout ratios were 35.5% and 63.6% for the Standard and Heavily-Prompted programs, respectively. These figures confirm that the Heavily-Prompted program was less effective because the prompts

undermined the contingency between the cue and the response.

While prompting techniques undoubtedly can be used to great advantage, this experiment has established one boundary condition for their use. Learning is reduced when the prompts are of such a nature that it is possible for the student to respond correctly without attending to the cues.

REFERENCES

ANDERSON, R. C., & FAUST, G. W. The effects of strong formal prompts in programed instruction. *American Educational Research Journal,* 1967, **4,** 345–352.

BROWN, R. W. Format location of programed instruction confirmations. *Journal of Programed Instruction,* 1966, **3,** 1–4.

COOK, J. O. 'Superstition' in the Skinnerian. *American Psychologist,* 1963, **18,** 516–518.

COOK, J. O., & SPITZER, M. H. Supplementary report: Prompting versus confirmation in paired associate learning. *Journal of Experimental Psychology,* 1960, **59,** 275–276.

HOLLAND, J. G. Response contingencies in teaching-machine programs. *Journal of Programed Instruction,* 1965, **3,** 1–8.

HOLLAND, J. G., & KEMP, F. D. A measure of programing in teaching-machine material. *Journal of Educational Psychology,* 1965, **56,** 264–269.

HOLLAND, J. G., & SKINNER, B. F. *The analysis of behavior.* New York: McGraw-Hill, 1961.

KEMP, F. D., & HOLLAND, J. G. Blackout ratio and overt responses in programed instruction: Resolution of disparate results. *Journal of Educational Psychology,* 1966, **57,** 109–114.

KRESS, G. C., JR., & GROPPER, G. L. A comparison of two strategies for individualizing fixed-paced programmed instruction. *American Educational Research Journal,* 1966, **3,** 273–280.

LEVINE, J. M. Prompting and confirmation as a function of the familiarity of stimulus materials. *Journal of Verbal Learning and Verbal Behavior,* 1965, **4,** 421–424.

SIDOWSKI, J. B., KOPSTEIN, F. F., & SHILLESTAD, I. J. Prompting and confirmation variables in verbal learning. *Psychological Reports,* 1961, **8,** 401–406.

WILLIAMS, J. P. Comparison of several response modes in a review program. *Journal of Educational Psychology,* 1963, **54,** 253–260.

WILLIAMS, J. P. Effectiveness of constructed-response and multiple-choice programing modes as a function of test mode. *Journal of Educational Psychology,* 1965, **56,** 111–117.

(Received March 14, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 2, 94–101

SEX AND SCHOLASTIC APTITUDE AS VARIABLES IN TEACHERS' RATINGS OF THE ADJUSTMENT AND CLASSROOM BEHAVIOR OF NEGRO AND OTHER SEVENTH-GRADE STUDENTS

LOIS-ELLIN DATTA, EARL SCHAEFER
National Institute of Mental Health

AND

MALCOLM DAVIS[1]
Washington School of Psychiatry

Both ethnic group and sex were significantly related to teachers' descriptions of the classroom behavior of 153 7th-grade students; girls and Other Ss were described more favorably than boys and Negroes on 75% of 64 behavioral rating scales. There were no interactions of ethnic group and sex. Comparisons of IQ equivalent subgroups showed, however, that the effect of ethnic group tended to be contingent on scholastic aptitude and was not dependent on sex; higher-IQ Negro Ss were described as favorably as were higher-IQ Other Ss but the lower-IQ Negro pupil was more likely than the lower-IQ Other child to be described as maladjusted, verbally aggressive, and low in task orientation. Boys, regardless of IQ or ethnic group, were described less favorably than girls.

Little is known about the associates of teachers' opinions of Negro and white pupils, although the importance of teachers' opinions and the fact that they are related to demographic variables such as sex and socioeconomic status appears to be well documented.

It has frequently been reported that teachers are more likely to describe boys than girls as maladjusted or as behavior problems (Beilin, 1959; Goldstein & Chorost, 1966; Long & Henderson, 1966; Vroegh & Handrich, 1966) and that children from well-to-do families are more likely than are lower-class children to meet with approbation and success in school (Charters, 1963). This latter finding has been interpreted as an indirect effect of social class differentials in academic preparation and opportunities (Sexton, 1961), but other studies have indicated that even among children of equal academic achievement who attend the same school, students whose parents are semi-skilled, unskilled, or unemployed are described less favorably than are students from upper- and middle-class families (Davidson & Lang, 1960).

There is considerable evidence that students who are described unfavorably by their teachers tend (*a*) to describe themselves unfavorably, (*b*) to be aware of their teachers' poor opinion of them, and (*c*) to receive lower grades than students whom the teacher describes favorably (Davidson & Lang, 1960; de Groat & Thompson, 1949; Fox, Lippitt, & Schmuck, 1964; Goldblatt & Tyson, 1962). Despite agreement that teacher attitudes toward Negro children should be highly important for their classroom behavior (Clark, 1963; Coleman, 1966; Deutsch, 1960; Katz, 1964; Riessman, 1962), Katz (1967) has concluded that there has been no adequate assessment of the attitudes of white teachers toward minority-group pupils.

The present analyses were undertaken to identify for use in planning an inservice teacher education program to facilitate integration, the extent to which the variables of sex and scholastic ability are associated with teachers' descriptions of Negro and Other-than-Negro students.

[1] The authors wish to thank Karen Pettigrew for statistical guidance and for computation of the tests for linearity of regression and Ann Drake for her assistance in the analysis and interpretation of results.

METHOD

Subjects

The subjects (*Ss*) were selected from seventh-grade classes in a northern Virginia suburban community which had integrated its schools in the year of the study. Ethnic group membership was inferred from attendance at segregated schools during the prior year. By this criterion, 9.8% (199) of all seventh-grade students were Negro and 90.2% (1931) were Other-than-Negro. (These students will hereafter be referred to as Others.[2]) The small total population of Negro students precluded selecting proportional random samples of Negro and Other *Ss*. In order to include as many Negro *Ss* as possible within the practical limits of teacher contact, the junior high schools with the largest numbers of Negro students were identified. Eighty-nine percent of the Negro students attended three of the system's six junior high schools. It was not possible to estimate social-class membership directly for each student; however, two of the schools were in neighborhoods judged as lower class (mixed residential and commercial buildings, low-cost housing, poor upkeep of buildings) and one, by the same criteria, was judged to be lower middle class. As these schools draw students from neighborhoods surrounding them, the social class of the student group may be at least generally inferred. In the two schools judged to draw lower-class students, 24.5% and 27.7% of the seventh-grade students were Negro. In the third school, which was judged to draw lower middle-class students, 9.8% of the seventh-grade students were Negro. A sample of 100 Negro and 100 Other *Ss* was chosen by a table of random numbers from the total of 177 Negro and 805 Other students attending these three schools. The students were not selected with consideration as to sex.

Procedure

One teacher was selected at random from each *S*'s schedule card. Both to conserve teacher time and to avoid, insofar as possible, biasing the data with respect to individual differences among raters, no teacher was asked to evaluate more than five students. In the "lower middle-class" school, all 17 seventh-grade teachers were included in the survey; in the other two schools, 76.7% (23 out of 34) of the seventh-grade teachers participated. Twenty of these 40 teachers taught general education courses, five taught mathematics, and the remaining 15 taught subjects such as physical education, language, music, and art. Thirty-six of the teachers were white and 30 were women. Class assignment in the system, and therefore selection for inclusion in the sample, was random with respect to pupil and teacher ethnic group and sex.

The teachers were contacted by mail. The covering letter stated that the intention of the study was to standardize the instruments and requested the teacher's cooperation in describing the adjustment and classroom behavior of the five students selected for him or her.

The questionnaires were distributed in November 1965. Of the 200 eligible students, 178 were rated and 153 records were sufficiently complete to be included in the analyses. There were significant differences in the proportion of records returned by ethnic group (92% returns for Other *Ss* as compared with 61% returns for Negro *Ss*), but not by sex. Nonreturns were due primarily to insufficient teacher time and to the fact that 11 pupils had left the school system between the time of selection and rating.

Measures

The teachers were asked to rate the students' adjustment on the following scale which was developed by Ullman (1952) and modified by Glidewell, Domkee, and Kantor (1963). The first two categories of Glidewell et al.'s scale were revised to emphasize social rather than academic accomplishment as the criterion of adjustment.

1. Well adjusted. A happy child who is well adjusted in his relationships with others and in his activities.
2. No significant problems. A child who gets along reasonably well and has little or no difficulty adjusting to others or to classroom activities.
3. Subclinically disturbed. A child who is not so happy as he might be; has moderate difficulties getting on; and to whom growing up represents something of a struggle.
4. Clinically disturbed. A child who has, or at his present rate is likely to have, serious problems of adjustment, and needs clinical help because of such problems.

It will be noted that with this instrument the teachers were not asked to identify students who presented problems in classroom management; attention was rather directed toward a more clinical definition of social and emotional adjustment.

The teachers were then asked to rate each of their students on the Classroom Behavior Inventory, (CBI), a recently developed 320-item questionnaire (Schaefer, Aaronson, & Burgoon, 1966). The questionnaire items were intended to describe behavior and to reduce as much as possible inferences about motives and feelings. Sample items included: "Often disagrees with what others suggest," "Brags how he is able to outwit others," "Begins work at once, as soon as something is assigned," "Seldom talks to other children before or after class," "Sticks to old ways of doing things; hates to make changes."

The teacher was asked to describe the behavior of each child for each item, with the following response options: 1. Not at all like the child, 2. Very

[2] The community, while predominantly white, includes a variety of ethnic groups. No direct information on family background can be obtained for the *Ss*. The students who had not attended a Negro school are therefore described as "Others" and the variable will be referred to as ethnic group rather than race.

little like the child, 3. Somewhat like the child, and 4. Very much like the child. The specific instructions were:

> Please give a response to every item and base your response upon your personal observation and experience with the pupil. In the case of items relating to behavior which you have not observed, respond as you would expect this child to behave as a general rule.

There are 64 five-item scales. Scale reliabilities for the sample of 153 *S*s as estimated by Kuder-Richardson formula 20 ranged from .73 to .96. The median internal consistency scale reliability was .86. A principal components analysis, Varimax rotation, yielded three factors. Scales describing perseverance, conscientiousness, concentration, achievement orientation, academic seriousness, and methodicalness had loadings of .76–.86 on Factor I, "Positive task orientation." Irritability, argumentativeness, attention seeking, boastfulness, quarrelsomeness, and dominance had loadings of .86–.93 on Factor II, "Verbal aggression." Active helpfulness, cheerfulness, and gregariousness had high negative loadings and social withdrawal, depression, and emotional passivity had high positive loadings on Factor III, "Introversion-extroversion." Adjustment ratings correlated .43, .43, and −.48, respectively, with the three factors. The average scores of each *S* were computed for the six scales with the highest loadings on each of the three factors. (Due to computer limitations, true factor scores could not be computed. The average scores would be expected, however, to correlate highly with the true factor scores.)

Scholastic aptitude was estimated by the California Mental Maturity Test, Short Form, (CMMT), which had been administered in the last part of the sixth grade. CMMT IQ equivalents were used in the analyses as measures of scholastic aptitude. Also reported are grade-equivalent scores for the Iowa Silent Reading Test, which was administered at the beginning of the seventh grade.

RESULTS

The distribution of adjustment ratings by sex and ethnic group is shown in Table 1. Only 4% of all *S*s were described as clinical problems, 33% of the students were described as very well adjusted, 40% were described as presenting no problems and 23% were described as subclinical problems, a distribution similar to that reported by Ullman (1952) for ninth-grade white students. Clinically and subclinically maladjusted categories were pooled for the following 2 × 3 comparisons. Girls were significantly more likely than boys to be rated as well adjusted ($\chi^2 = 15.24$, $p < .01$), whether the students were Other ($\chi^2 =$ 6.43, $p < .05$) or Negro ($\chi^2 = 12.64$, $p < .01$). Other students were more likely than Negro students to be rated as well adjusted if the students were girls ($\chi^2 = 8.40$, $p < .05$), but not if the students were boys ($\chi^2 = 2.30$, $p > .50$).

TABLE 1

PERCENTAGES OF OTHER AND NEGRO, MALE AND FEMALE SEVENTH-GRADE STUDENTS RATED AS ADJUSTED AND MALADJUSTED BY THEIR TEACHERS

Race and sex	N	Teacher ratings of adjustment: Very well	No problems	Subclinical	Clinical
Other					
Male	42	31%	33%	33%	3%
Female	47	53%	32%	13%	2%
Negro					
Male	31	16%	35%	35%	13%
Female	29	24%	66%	10%	0%
Total	149	33%	40%	23%	4%

Among Other students there was no significant relation between IQ and adjustment ($r = -.05$, $p > .10$), whether the students were girls ($r = -.16$, $p > .10$) or boys ($r = .06$, $p > .10$). There was no relation between scholastic aptitude and adjustment for Negro girls ($r = -.01$, $p > .10$) but the less able Negro boys were more likely than were the brighter Negro boys to be described as subclinically or clinically disturbed ($r = -.48$, $p < .01$). Scholastic aptitude is thus shown to be a significant associate of teachers' ratings of adjustment only with Negro boys.

Analyses by Ethnic Group and Sex for the Whole Sample

A two-way unweighted means analysis of variance (Winer, 1962) was computed for ethnic group and sex for each of the 64 scales. On 48 of the 64 scales, the effect of sex was significant at the .05 level. On 46 scales, the effect of ethnic group was significant at the .05 level. On only one scale, "Work fluctuation," there was a significant Ethnic Group × Sex interaction. The mean IQ for girls, 101.4, did not differ significantly from the mean IQ for boys,

100.2; the mean IQ for Negro *Ss*, 87.5, was significantly lower than the mean IQ, 113.0, for Other *Ss*. While scholastic aptitude would not seem to account for the differential description of boys and girls, the characteristics attributed to Negro *Ss* as compared to Other *Ss* might be associated with lower scholastic ability rather than with ethnic group per se.

There are considerable methodological difficulties in isolating the variance due to ethnic group and sex from that due to scholastic aptitude because of the low overlap in the CMMT distributions and the asymmetry of IQ/adjustment and IQ/CBI scale correlations. As an example, "submissive" correlated −.30 with CMMT IQ for Other girls and +.40 for Negro boys; "methodical" correlated .51 and .40 with IQ for Negro girls and Negro boys, but for Other boys, $r = .01$. Analysis of covariance was not appropriate since, except for Negro *Ss*, there was no reliable evidence of a linear relation between the dependent variables and scholastic aptitude as measured by the CMMT. Despite restrictions on generalizations to the upper end of the Other IQ distribution, the most defensible approach seemed to be a three-way analysis of variance for ethnic group, sex, and IQ.

Analyses by Sex, Ethnic Group, and Scholastic Aptitude for IQ Equivalent Subsamples

The overlap between Other and Negro CMMT distributions ranged from IQ 68 to 114. Other students with IQs above 115 ($N = 37$) were dropped from these analyses in order to facilitate matching IQ groups; all *Ss* with IQs below 68 ($N = 6$) were also dropped. "Higher" *Ss* were defined by IQs between 99 and 114; "lower" *Ss* were defined by IQs between 68 and 98. Two-way unweighted means analyses of variance (ethnic group and scholastic ability) were computed for each of the 64 scales for the IQ-selected samples of boys and girls. Of the 128 *F*'s, 14 were significant for ethnic group at ≤.10 level, 21 were significant for scholastic aptitude, and 27 of the interactions were statistically

TABLE 2

MEANS AND *N*s OF TEACHER RATINGS ON FACTOR SCORES: I, TASK ORIENTATION (TO); II, VERBAL AGGRESSION (VA); EXTROVERSION (E) AND INTROVERSION (I) BY SCHOLASTIC APTITUDE, ETHNIC GROUP AND SEX FOR SEVENTH-GRADE STUDENTS

Group[a]	*N*	Factor score means: Mean CMMT IQ	Mean reading achievement	TO	VA	E	I
Males							
Negro							
Low IQ	15	81.5	3.9	9.8	12.1	11.6	12.7
High IQ	11	105.2	4.9	11.7	9.9	13.8	12.1
Other							
Low IQ	8	78.9	4.5	13.5	10.1	14.2	11.3
High IQ	11	109.4	6.6	11.8	11.7	13.4	10.8
VH IQ	20	122.9	9.5	13.2	9.5	13.8	12.6
Females							
Negro							
Low IQ	18	86.7	5.4	11.7	9.1	13.8	10.5
High IQ	7	104.1	6.9	14.7	6.3	13.5	9.3
Other							
Low IQ	11	87.9	5.7	15.2	6.8	14.4	10.3
High IQ	17	108.2	7.6	15.7	7.1	15.5	8.9
VH IQ	17	123.5	9.6	15.0	8.1	15.5	9.1

[a] CMMT IQ (very high, 115–140; high, 99–114; low, 68–98).

reliable. For expository simplicity we will describe the results primarily in terms of the 2 × 2 × 2 unweighted means analysis of variance completed for average scores. To facilitate interpretation of the introversion-extroversion scores, the scales which measured introversion and extroversion were considered separately. The means for the eight cells for each of the four average scores are shown in Table 2.

Sex. As Table 3 indicates, on three of the four average scores, sex accounted for a significant portion of the variance: girls tend to be rated higher than boys in task orientation and were less likely than boys to be described as either verbally aggressive or as high in introversion; boys, regardless of ethnic group, were more likely than girls to be described as withdrawn, asocial, and emotionally passive. To see if the effect of sex was linear, analyses of variance were computed on factor scores for the Other students only by IQ (highest, higher, and lower) and sex: again, regardless of IQ, boys were described as more withdrawn, asocial, and emotionally passive than were girls. For the Other students only, with the highest IQ *Ss* in-

TABLE 3

F RATIOS FOR UNWEIGHTED MEANS ANALYSES OF VARIANCE BY SCHOLASTIC APTITUDE, ETHNIC GROUP, AND SEX COMPUTED FOR AVERAGE SCORES[a] DERIVED FROM TEACHERS' RATINGS OF THE CLASSROOM BEHAVIOR OF SEVENTH-GRADE STUDENTS

Source	*F* ratios for factor scores			
	TO	VA	E	I
A Scholastic aptitude	1.77	1.13	0.81	2.48
B Ethnic group	8.91***	0.31	3.81*	1.98
C Sex	14.27****	24.77****	2.92*	11.33***
A × B	4.82**	5.61**	0.42	.00
A × C	1.42	0.43	.00	.41
B × C	0.83	0.20	.00	.80
A × B × C	1.44	0.07	3.20*	.03

[a] I, Task Orientation (TO); II, Verbal Aggression (VA); III, Extroversion-Introversion (E-I); and Extroversion (E) and Introversion (I).

* $p < .10$.

** $p < .05$.

*** $p < .01$.

**** $p < .001$.

cluded, girls were significantly more often described as cheerful, gregarious, and actively helpful. Sex did not interact with ethnic group or scholastic aptitude on any factor except extroversion (at the .10 level).

Scholastic aptitude. Scholastic aptitude did not account for a significant portion of the variance on any factor.

Ethnic group. Ethnic group accounted for a significant portion of the variance on task orientation and extroversion. Negro students were more likely than were Other students to be described as low in task orientation and, at the .10 level, were less likely to be described as helpful, cheerful, and gregarious. For two factors, task orientation and verbal aggression, the interaction of ethnic group with scholastic aptitude was significant. On task orientation, the difference between Negro and Other students was greater among lower-IQ than among higher-IQ *Ss*, regardless of sex. On verbal aggression, the effect of ethnic group depended on scholastic aptitude: among brighter students, Negro *Ss* were described as less verbally aggressive than were Other *Ss*, while among low-IQ students, Negro *Ss* were described as more verbally aggressive than were Other *Ss*. The interaction of sex, ethnic group, and scholastic aptitude was significant at the .10 level for extroversion; the difference between boys and girls, and Other and Negro students depended mainly on the low average rating of 11.6 received by low-IQ Negro boys and the high ratings of 15.5 and 15.5 received by high- and very high-IQ Other girls.

The results of these analyses and of analyses of the 64 individual scales for the IQ equivalent subsamples are summarized in the following section.

1. Among Negro students, low-IQ *Ss* were more frequently described as low in adjustment, low in task-orientation, verbally aggressive, rebellious, asocial, and unruly. They were seen neither as well behaved nor as studious. Higher-IQ Negro *Ss* were likely to be described as task oriented, methodical, persevering, sociable, trustful, submissive, and as low in rebelliousness and verbal aggression. They were seen as well behaved, hardworking, studious pupils.

2. Among Other students, low-IQ *Ss* were more frequently described as task oriented, low in verbal aggression, compliant, cooperative, and considerate. The CMMT and reading achievement scores suggest that they are not achieving despite the high level of academic and social effort described: not unpredictably, low-IQ Other *Ss* were also seen as lower in self-esteem, and as more tense and fatigued than were the other three subgroups. The higher IQ Other *Ss* were likely to be described as task oriented, verbally aggressive, inquisitive, enthusiastic, and as leaders.

In the IQ equivalent subsample, then, ethnic group is associated with descriptions of classroom behavior, but the direction of the association tends to be contingent on the scholastic aptitude of *S* and is due primarily to the unfavorable descriptions of Negro as contrasted to Other low-IQ students, particularly Negro boys. The classroom behaviors described by the teachers suggest that the low-IQ Other student and the higher IQ Negro student tend to cope with the demands of school by working hard and carefully and by offering few problems in classroom manage-

ment. The low-IQ Negro student appears to have resigned from the educational process and is seen as poorly adjusted, unruly, and uninvolved, behaving in ways generally considered to be associated with educational failure and classroom management difficulty. The data suggest that the Other, higher IQ student may be freer to adopt a mode of response characterized by exploration, dominance, independence, and academic interest.

3. Regardless of ethnic group or scholastic aptitude, boys were more likely to be described as hyperactive, asocial, verbally and physically aggressive, and tense and were less likely to be described as friendly, methodical, persevering, task oriented, and well adjusted than were girls. They were not likely, however, to be described as lower in such traits as enthusiasm, inquisitiveness, leadership, verbal expressiveness, academic ability, nor as higher in conformance.

DISCUSSION

One important finding of this study is that the unfavorable description of the Negro student is associated primarily with *S*s of lower scholastic aptitude. At least two questions may be raised concerning the interpretation of this finding, in addition to the limitations imposed by the restricted range of scholastic aptitude and by the small *N*s: (*a*) is it an artifact of social class and (*b*) are teachers' descriptions observations of actual behavior or perceptions that would be considered biased or limited in comparison to what other observers might report?

1. We do not know whether the interaction among scholastic aptitude and ethnic group is associated with Negroes in particular, with minority groups, or more generally with social class. The description of the low-IQ Other student suggests some effort by the child to conform to the demands of middle-class parents for good grades and good behavior in school while the description of the low-IQ Negro student does not and seems more consistent with behavior generally attributed to children from low-income families.

Schmuck and Luszki (1966) have reported that in a small, midwestern community, there were no differences in achievement, self-ratings, and teachers' classroom behavior ratings when socioeconomic status was carefully matched for Negro and white students. They conclude that relations among race, self-esteem, and achievement are confounded in other studies with social class. Only 63 pairs of students, ranging in age from 8 to 16 years, were involved in the study; a larger sample may be needed at each age and grade level to test the social-class interpretation rigorously and the nature of the community might itself be a relevant variable (cf. Davidson & Lang, 1960). In our sample, it is possible that despite the somewhat homogeneous neighborhoods, Other and higher-aptitude Negro students came from less-deprived homes than did lower-aptitude Negro *S*s.

2. We have referred to teachers' descriptions rather than to either teachers' perceptions or students' behavior. The teachers had access to intelligence test and reading scores and knew the students' ethnic group and sex. Whether in this instance the frequently postulated interaction between expectations and observations is weighted more heavily with expectations or was formed by observation relatively independent of teachers' a priori values, is moot.

Considering the "observation" interpretation, results similar to ours have been noted for younger children whose teachers had volunteered for the assignment. Such teachers might be expected to be somewhat more favorably disposed toward the children than public school teachers assigned to schools in low-income Negro neighborhoods. Lamb, Ziller, and Maloney (1965) found, for example, that white girls gained most from Headstart experiences and that the Negro boy was both least favorably described by his teacher and least likely to benefit from the preschool program. The description of the brighter Negro students as more "compliant" is congruent with the report that in comparison to white liberal-arts college students, Negro liberal-arts

students scored higher on deference and lower on exhibitionism, autonomy, and dominance on the self-descriptive Edwards Personal Preference Schedule (Pettigrew, 1964).

Considering the "opinion" interpretation, Rosenthal (1966) and Flowers (1966) have shown that students' IQs and classroom performance tend to increase when teachers are led to believe the child's intellectual potential is high relative to an equally bright control *S* whom the teacher believes has lower intellectual potential. Their effects have been demonstrated, however, more reliably in younger than older children. Rotter's (1967) studies also suggest that "....preconceptions influence one's perceptions and evaluations and that these might lead to differential treatment." Groups of white teachers read vignettes reporting a child's behavior. Analyses are reported for vignettes which differed in the sex ascribed to the child and classroom behavior: e.g., orderly/disorderly, "Ann carried a classmate's books"; "Billy carried a classmate's books." Sex and the interaction of sex with classroom orderliness were significant associates of teacher's rating on many of 80 bipolar scales, including a rating of boys as "dirtier." Race (white and Negro) and social class (middle and low) were also variables in the study and the technique should be of considerable value in separating "opinion" and "observations" in the descriptions of minority children by majority teachers.

Davidson and Lang's comments (1960, p. 114) on the antecedents of scholastic difficulty may be relevant here:

> It is likely, therefore, that a lower class child, especially if he is not doing well in school, will have a negative perception of his teachers' feelings toward him. These negative perceptions will in turn tend to lower his efforts to achieve in school and/or increase the probability that he will misbehave. His poor school achievement will aggravate the negative attitudes of his teachers toward him, which will in turn affect his self-confidence, and so on. This vicious entanglement must be interrupted at some point. The point of attack may well be the teacher whose capacity to reflect feelings conducive to the child's growth should be of concern to educators.

To this we would add that both the relatively high correlation between IQ and adjustment for Negro males and the analysis of variance results suggest that the low-IQ Negro is alienated from the school situation, that is, is not task oriented and is verbally aggressive and withdrawn. This finding supports the need for programs designed to raise the level of intellectual performance before the vicious cycle of low achievement, teacher rejection, and child alienation begins.

A second major finding in this study is that boys were described as less task oriented, more verbally aggressive, and more introverted than girls. The latter result is unexpected as girls have been generally described as less outgoing and more introverted than boys. Our initial interpretation of an introversion-extroversion analysis was that helpful classroom behavior, "extroversion," might be accounting for most of the variance: clearer sex differences were found, however, for the "introversion" than for the "extroversion" factor scores.

Interpretations of the sex differences in personality and attainment have ranged from biological forces to a greater disparity for boys than for girls between the classroom demands of female teachers and socially defined behavior appropriate to the student's sex. Maccoby (1966), reviewing this literature, has noted that peer group pressures on boys are often directed to nonacademic pursuits;

> that boys are more frequently engaged in efforts to achieve autonomy, especially in relation to their mothers, with the result that they are less willing to accede to the demands of their predominantly female teachers; and that even in high school, boys are more likely to do poorly in subjects that bore them [p. 32].

The observed higher "introversion" ratings for boys may thus indicate apathy in a traditionally academically oriented classroom situation rather than a more general trait.

REFERENCES

BEILIN, H. Teachers' and clinicians' attitudes toward the behavior problems of children: A reappraisal. *Child Development*, 1959, **30**, 9-25.

CHARTERS, W. W., JR. The social background of

teaching. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally, 1963. Pp. 715–813.

CLARK, K. Educational stimulation of racially disadvantaged children. In A. H. Passow (Ed.), *Education in depressed areas*. New York: Teachers College Press, 1963. Pp. 142–162.

COLEMAN, J. S.. et al. Equality of educational opportunity. Report No. OD-38001, 1966, United States Office of Education, Washington, D. C.

DAVIDSON, H., & LANG, G. Children's perceptions of their teachers' feelings toward them related to self-perception, school achievement and behavior. *Journal of Experimental Education*, 1960, **29**, 107–118.

DE GROAT, A. F., & THOMPSON, G. G. A study of the distribution of teacher approval and disapproval among sixth grade pupils. *Journal of Experimental Education*, 1949, **18**, 57–75.

DEUTSCH, M. *Minority group and class status as related to social and personality factors in scholastic achievement*. Monograph No. 2. Ithaca: The Society for Applied Anthropology, 1960.

FLOWERS, C. E. Effects of an arbitrary accelerated group placement on the tested academic achievement of educationally disadvantaged students. Paper presented at the meeting of the American Psychological Association, New York, September 1966.

FOX, R. S., LIPPITT, R. O., & SCHMUCK, R. A. Pupil-teacher adjustment and mutual adaptation in creating classroom learning environments. Intercenter Program of Studies on Children, Youth and Family Life. Document Series 16, Cooperative Research Project No. 1167, Institute for Social Research, University of Michigan, 1964.

GLIDEWELL, J. C., DOMKE, H. R., & KANTOR, M. Screening in schools for behavior disorders: Use of mothers' report of symptoms. *Journal of Educational Research*, 1963, **56**, 508–515.

GOLDBLATT, H., & TYSON, C. An ethnic study: Pupil evaluations. Research Library Report No. 12, 1962, Commission on Human Rights, New York.

GOLDSTEIN, K. M., & CHOROST, S. B. Preschool and background factors in the school adjustment of culturally disadvantaged children. Paper read at the meeting of the American Psychological Association, New York, September 1966.

KATZ, I. Review of evidence relating to effects of desegregation on the intellectual performance of Negroes. *American Psychologist*, 1964, **19**, 381–399.

KATZ, I. Some motivational determinants of racial differences in intellectual achievement. *International Journal of Psychology*, 1967, **2**, 1–78.

LAMB, H. E., ZILLER, R. C., & MALONEY, A. W. The development of self-other relationships during Project Headstart. Project No. OEO-511, 1965, University of Delaware. (Mimeo)

LONG, B., & HENDERSON, E. H. Self-social concepts of disadvantaged school beginners. Paper read at the meeting of the American Psychological Association, New York, September 1966.

MACCOBY, E. E. Sex differences in intellectual functioning. In E. E. Maccoby (Ed.), *The Development of sex differences*. Stanford: Stanford University Press, 1966. Pp. 25–55.

PETTIGREW, T. *A profile of the Negro American*. New Jersey: Van Nostrand, 1964.

RIESSMAN, F. *The culturally deprived child*. New York: Harper and Row, 1962.

ROSENTHAL, R. Teachers' expectancies: Determinants of changing children's IQ by changing teachers' expectations. Paper read at the meeting of the American Psychological Association, New York, September 1966.

ROTTER, G. S. The effect of sex identification in teacher evaluation of pupils. Paper read at the meeting of the Eastern Psychological Association, Boston, April 1967.

SCHAEFER, E. S., AARONSON, M., & BURGOON, B. Classroom Behavior Inventory (Form for Grades 3 through 12). National Institute of Mental Health, 1966. (Mimeo)

SCHMUCK, R., & LUSZKI, M. A comparison of Negro and white students in several small midwest communities. Paper read at the meeting of the American Psychological Association, New York, September 1966.

SEXTON, P. *Education and income: Inequalities in our public schools*. New York: The Viking Press, 1961.

ULLMAN, C. A. *Identification of maladjusted children*. Public Health Monographs, No. 7, 1952.

VROEGH, K., & HANDRICH, M. Sex role typing in the preschool years: An overview. Research Report No. 13, Vol. 3. Chicago: Institute for Juvenile Research. 1966.

WINER, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

(Received March 27, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 2, 102–110

LEARNING ABILITIES OF NORMAL AND RETARDED CHILDREN AS A FUNCTION OF SOCIAL CLASS[1]

JACQUELINE L. RAPIER[2]
University of California, Berkeley

This study compared the learning ability of normal and retarded elementary school children (N = 80) from high- and low-SES backgrounds on a series of learning tasks. On the first day, all Ss learned a serial list and a paired-associate list. 24 hr. later, Ss were divided into an experimental and a control group on the basis of CA, IQ, and SES. The experimental group learned a 2nd list of paired-associates, under conditions of mediation, that is, sentences were provided linking the pairs on the 1st trial. The control group learned paired associates without instruction in mediation. 1 wk. later, all Ss learned a 3rd list of paired associates. Results showed IQ differences in performance in both SES groups on serial and paired-associate learning. A significant mediation effect was found on the 2nd day, but this effect did not transfer to the learning of paired associates 1 wk. later in any group. However, over the 3 lists of paired associates, an increasing superiority in performance was found for the low SES retardates as compared to the high-SES retardates.

Sarason and Gladwin (1958) note that a large number of retardates are from the most economically and socially underprivileged families in society and suggest that the impoverished environment has provided minimal opportunity for the learning of skills which are tapped by current intelligence tests. Since these retardates usually are not identified until they enter school and show difficulty in dealing with verbal tasks, Sarason hypothesizes that language skills receive little stimulation in their home environment. Diagnosis is complicated by the fact that retardates are usually most deficient in the verbal area, so it has become a problem to differentiate the organically retarded child (organic retardation) from the child who appears retarded due to an early history of environmental deprivation (cultural retardation).

Jensen (1967) suggests the use of a variety of direct learning tests to assess the child's disability rather than employing standardized measures of past achievement. A series of experiments by Jensen and Rohwer (1963a, 1963b, 1965) showed that serial and paired-associate learning represent quite different levels of learning. Although both tasks involve rote learning skills, paired-associate learning appears to be more complex due to the more important role of verbal mediation. Experiments which directly instructed subjects (*S*s) to use verbal mediators showed greatly facilitated paired-associate learning, but verbal mediators did not influence the rate of serial learning. Among normal children, ability in paired-associate learning increases with age up to 18 years, presumably due to the increasing use of verbal facilitative devices. There appears to be no similar increase in the ability to learn a serial list beyond age 8, which Jensen suggests is related to its lack of dependence on verbal mediation. On the basis of this research, Jensen proposes that serial learning more nearly measures learning ability relatively unaffected by *S*'s previous verbal experiences while paired-associate learning is dependent on the richness of *S*'s verbal experiences and on the availability of relevant verbal mediators.

Various investigators (Griffith, Spitz, & Lipman, 1959; Jensen, 1965; Rieber, 1964) agree that retardates do not employ ver-

[1] This report is based on a doctoral dissertation submitted to the University of California, Berkeley.

[2] Now with the Palo Alto Unified School District, Palo Alto, California. The author is indebted to the dissertation committee, particularly the Chairman, Arthur R. Jensen and to a number of school districts in the San Francisco Bay Area who cooperated in the study.

bal mediators as effectively as normals. From an educational point of view, it is important to know if the retarded child is unable to elicit verbal associations as mediators because he has not had the opportunity to form necessary associations prior to the learning situation or because he has a neurological deficit in association formation. Obviously, the two types of slow learners should have quite different types of educational treatment as the former may be able to overcome his handicap through remedial training while the latter requires more appropriate educational goals.

Jensen (1967) suggests that the culturally retarded child will learn a serial list more efficiently than a paired-associate list as serial learning does not depend as much on the richness of the child's early language experience. The organically retarded child will be slow in both serial and paired-associate learning even if his environment has been good, because of a basic deficiency in the neural equipment for learning.

The present experiment was designed to explore the relation of serial and paired-associate learning to IQ and socioeconomic status (SES). The specific hypotheses are as follows:

1. Since it is hypothesized that serial learning is relatively unaffected by verbal mediation but does reflect learning ability, IQ will predict serial learning ability more effectively in the high- than in the low-SES groups.

2. Paired-associate learning, which is believed to involve verbal mediational processes to a greater extent than serial learning, will closely reflect IQ differences in both high- and low-SES groups.

3. The normal and retarded groups from high- and low-SES will show increased speed of paired-associate learning after verbal mediation instruction as compared with control groups not receiving instruction.

4. Transfer of the mediation technique to a new paired-associate task 1 week later should be greater for the normal than for the retarded groups in both SES groups. However, the retarded low-SES group will show greater transfer of mediation tendencies than the retarded high-SES group.

METHOD

Subjects

Eighty white American *S*s were selected from public elementary schools in Alameda and Contra Costa County, California. The age range was 91 to 154 months. The *S*s were equally divided between two socioeconomic classes, an upper and a lower class. The *S*s were classified as high SES if their father was in a professional, semiprofessional or managerial occupation and had completed 2–4 years of college. The *S*s were placed in the low-SES group if their father was engaged in unskilled or semiskilled work and had not gone beyond high school. There were two IQ levels in each SES group: a normal IQ group and a retarded IQ group. The normal group was selected on two bases: an IQ of 100–110 on a group test, the Kuhlman-Anderson Intelligence Test, and average achievement for grade placement on the Stanford Achievement Test. The retarded group was selected on two bases: an IQ score of 63–78 on a recent (not more than 2 years) Stanford-Binet Intelligence Scale and placement in a special class for educable mentally retarded children. It was not possible to obtain individual test scores on the normal group. The Kuhlman-Anderson is reported to measure substantially the same thing as the Stanford-Binet (Cronbach, 1960). Furthermore, it is rare that a child who receives an average IQ score on a group test and achieves at grade level would score below average on an individual test.

Since the study rests on the hypothesized relationship of verbal ability to SES, additional information was sought on *S*'s verbal development. Vocabulary is a widely used measure of language development, and it is susceptible to environmental stimulation (McCarthy, 1954). The Peabody Picture Vocabulary Test (Dunn, 1965) is an individual picture vocabulary test where *S* is required to choose the picture on the plate which best illustrates the meaning of the word provided by the experimenter (*E*). The Peabody seemed preferable to the standard vocabulary test where *S* must define the meaning of the word since the latter depends on *S*'s ability to express his ideas verbally. Retarded *S*s may comprehend the meaning of the word, but lack the fluency of speech to provide a satisfactory definition.

The Peabody was administered to all *S*s. The high-SES group had a mean raw score of 78.55 and a mean MA of 127.65 months. The low-SES group had a mean raw score of 72.37 and a mean MA of 112.37 months. The MA difference between SES groups is significant ($F = 7.18$, $df = 1/72$, $p < .01$), indicating that SES is related to verbal ability in this population.

The *S*s were assigned to one of the four groups on the basis of their age, IQ, and SES. There were

TABLE 1

CHARACTERISTICS OF THE FOUR GROUPS

Group	Measure	CA (months)	MA (months)	IQ
High-SES normal	Mean	124.40	130.75	105.10
	Range	94–147	99–152	100–110
	SD	18.48	19.87	3.70
Low-SES normal	Mean	126.10	131.55	104.50
	Range	96–149	103–155	100–110
	SD	16.63	16.29	3.23
High-SES retarded	Mean	123.95	88.60	71.45
	Range	92–154	60–115	63–78
	SD	19.95	15.48	4.95
Low-SES retarded	Mean	123.90	86.60	70.20
	Range	91–150	68–103	63–78
	SD	17.21	10.25	3.64

20 *S*s in each group. The ratio of boys to girls was kept approximately equal in each of the four groups, 15 boys to 5 girls. Children whose health records indicated sensorimotor disabilities or emotional disorders were excluded from the study. Table 1 presents information on the characteristics of each group with the means, standard deviations, and range for CA, MA, and IQ for the 80 *S*s.

When attrition occurred within one of the four groups, an additional *S* was drawn at random from a pool of *S*s and assigned to that particular group. Four *S*s from the high-SES retarded group and two from the low-SES retarded were eliminated due to failure to reach the criterion on original learning. In addition, five *S*s were replaced when they were absent on the second day or final week of the experiment.

Design

Day 1. All *S*s were given two tasks, serial and paired-associate learning. In serial learning, a different order of pictures was used for each *S* in a group but the same 20 different orders were repeated in each group. In paired-associate learning, the position of the pairs was randomly changed from trial to trial by shuffling the cards between trials. In order to avoid practice and fatigue effects, half the *S*s learned the serial task first and then the paired associates; the other half learned the paired-associate task followed by the serial list.

Day 2. The second testing session occurred 24 hours later. The *S*s were assigned to an experimental group and a control group on the basis of their CA, IQ, and SES. The experimental and control groups learned a new list of paired associates under different conditions of instruction. The experimental group received instruction in the use of verbal mediators (mediation) while the control group did not receive such instruction (nonmediation). The paired-associate list presented on Day 1 was counterbalanced with the paired-associate list presented on Day 2 since one could not safely assume the two lists were equivalent in difficulty.

Day 3 (1 week later). All *S*s learned a third list of paired-associates under the same nonmediation conditions.

Stimulus Materials

The stimulus material of both the serial and paired-associate tasks consisted of black and white pictures of common objects, for example, ball, house, table. Pictures were cut from preprimer workbooks and mounted on gray cardboard. There were four sets of pictures, each set using different pictures. One set used for serial learning consisted of nine pictures, each mounted on 4 × 4-inch cardboard. Three sets used for paired-associate learning consisted of nine pairs of pictures on 5 × 7-inch cards. On one side a single picture appeared, and on the other side of the card, the same picture was paired with another picture unrelated to the stimulus picture. Paired associates were made from pictures paired at random, but obvious relation of sound and meaning were avoided between members of each pair.

Method of Stimulus Presentation

The *S*s were tested individually in an unused room in the school building. Testing conditions were reasonably comparable in terms of extraneous stimuli. The *S* was seen at about the same time for each session.

During the experiment, *E* sat at a table facing *S*. On Day 1, *E* gave the following instructions for serial learning:

> We're going to play a short game. See these cards in a row? When I turn the card over, there is a picture on the other side. Name each picture as I show it to you. [The *S* named each picture as *E* turned over the card.] Now, I want you to learn the names of all the pictures in this row. When I point to the card, you tell me what you think is on the other side. Then I will turn the card over and you can see if you are right.

Immediately following serial learning, *E* presented the paired-associate task.

> I'm going to show pictures like this (*E* showed one stimulus picture on a sample card) then I'll turn the card over like this, and you'll see the same picture with another one next to it like this. I want you to say the name of the picture next to the first one. Let's see how fast you can learn which pictures go together. [The *E* presented the series of nine cards and *S* was asked to name the stimulus and response picture in each case.]

On subsequent trials, *S* was only required to name the response picture when he saw the stimulus picture. Whether *S* made a correct response or not, the card was turned over so that *S* could see the stimulus and response pictures side by side.

On Day 2, two conditions of instruction were used—mediation and nonmediation. In the mediation condition, a standard set of sentences linking the stimulus with the response was provided by *E* upon initial presentation of each pair. Providing a standard set of sentences seemed preferable to *S* making up his own sentences due to wide individual differences in the ability among *S*s to invent their own phrases. The *S* was asked to repeat the sentence after *E*. After the first trial, the procedure was essentially the same as in the previous day, learning of paired associates, that is, *S* had to anticipate the response term when the stimulus was shown. The *S* was discouraged from repeating the sentences after the first trial.

The standard set of sentences given by *E* for the first set of paired-associates were: (*a*) The SPOON falls out of the NEST. (*b*) I stuck the FLAG inside my SHOE. (*c*) The COMB dropped under the CHAIR. (*d*) I carried the BASKET inside the HOUSE. (*e*) My HAND winds the CLOCK. (*f*)The SCISSORS cut the LEAF. (*g*) The IRON takes a ride on the BICYCLE. (*h*) The TREE grows inside the CUP. (*i*) The GLASSES are eaten by the FISH.

The phrases for the second set were: (*a*) The CAR runs over the BALL. (*b*) The DESK hides some MONEY. (*c*) The FIRE burns the SAW. (*d*) The PIANO is behind the COAT. (*e*) The KEY locks the WAGON. (*f*) The BOAT scares the CAT away. (*g*) The BOX holds a TELEPHONE. (*h*) The CANDLES ride on the HORSE. (*i*) The BUS breaks the RING.

In the nonmediation condition of paired-associate learning, *E* asked *S* to name the stimulus and response terms on the first trial. After this, the procedure was the same as in the mediation condition.

On Day 3, the third list of paired associates was presented to all *S*s using the same procedure as in the first presentation on Day 1. At the conclusion of this session, all *S*s were asked if they had used any special method to help them learn the list of paired associates.

Criterion of learning on all four tasks was eight out of nine correct responses on any one trial. All tasks were *S* paced. The *S* was dropped from the experiment if he failed to reach the criterion within 15 trials.

RESULTS

The mean trials to criterion and standard deviations of the distribution for the four groups on Day 1 are presented in Table 2.

Individual comparisons among trial means were made by the Scheffé method (Hays, 1963). The statistical results support the following conclusions: (*a*) SES has no significant effect ($F = 1.08$, $df = 1/79$, $p > .10$) upon the learning of the two tasks; (*b*) IQ had a significant effect ($F = 16.00$, $df = 1/79$, $p < .001$) upon the learning of both tasks; (*c*) normal *S*s learned the serial list in fewer trials than the retarded *S*s ($F = 5.41$, $df = 1/72$, $p < .05$); (*d*) normal *S*s learned the paired-associate list in fewer trials than the retarded *S*s ($F = 5.41$, $df = 1/72$, $p < .05$); (*e*) in both IQ groups, normal and retarded, the paired-associate task was much more difficult to learn than the corresponding serial task ($F = 144.42$, $df = 1/72$, $p < .001$); (*f*) the order of presentation of the two tasks made no difference in the learning of the tasks ($F = 3.67$, $df = 1/79, p > .05$).

Table 2 also shows the means and standard deviations for the mediation and nonmediation groups on Day 2. It is obvious from the table that the effect of mediation was to reduce drastically the number of trials to criterion. Learning in the mediation groups was four times as fast as in the nonmediation group. Use of the Cochran test for homogeneity of variance revealed that the assumption of homogeneity was not tenable. Siegel (1956) recommends a nonparametric test, the Mann-Whitney U test, as the most useful alternative to the usual t test. Aside from the significant effect of mediation instruction, none of the other main effects was significant. In order to keep the overall error rate for two-way interactions in the range of .08, it was decided to conduct each separate test at .02. There were two significant interactions, IQ × Instructions and SES × IQ. Although both ability groups profited from the mediating instructions, they did not do so to the same degree. Normal groups surpassed retardates in speed of learning under mediated conditions. Normals maintained their superiority in nonmediated conditions. Inspection of the SES × IQ interaction shown in Figure 1 revealed that IQ differences in performance were solely a function of SES. Normals and retardates in high SES showed significant differences in performance, but this did not occur in the low-SES group where the normals and retardates learned at about the same rate.

There was one significant three-way interaction—SES × IQ × Treatment in the original analysis of covariance. It was

TABLE 2

MEANS AND STANDARD DEVIATIONS FOR TRIALS ON DAY 1, DAY 2, AND DAY 3

Group	Day 1				Day 2				Day 3					
	Serial learning		Paired-associate learning		Mediation		Nonmediation		Mediation		Nonmediation		Total	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
High-SES normal	3.85	1.46	6.75	3.24	1.50	.70	6.50	3.10	5.00	3.05	5.20	2.69	5.10	2.80
Low-SES normal	4.25	1.25	8.85	3.60	1.70	.67	7.40	2.59	5.20	1.75	6.70	2.26	5.95	2.11
High-SES retarded	5.10	1.55	10.55	3.42	3.30	1.49	12.10	3.92	10.40	3.92	9.70	3.23	10.05	3.51
Low-SES retarded	4.85	1.75	10.35	3.77	2.30	1.56	7.50	3.40	7.00	3.46	6.20	2.20	6.60	2.85

difficult to analyze the significance of this interaction by means of the Mann-Whitney U test due to the large number of separate tests involved. When such a large number of separate tests is conducted, there is always the possibility of some being significant by chance alone. Of the 28 tests, 20 had $p < .01$ which is more than one would expect from chance. The SES × IQ × Treatment can best be shown graphically by Figure 2. Examination of Figure 2 immediately points up the striking difference in performance between the two retarded groups. The high-SES retarded is the least efficient learner in both conditions, whereas the low-SES retarded does not differ appreciably from the two normal groups.

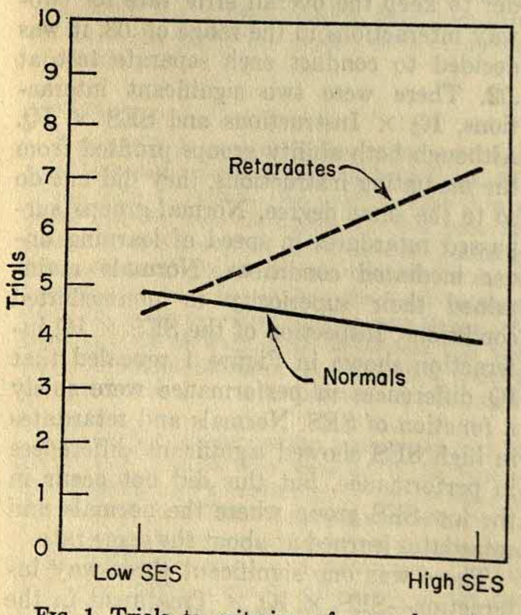


FIG. 1. Trials to criterion of normals and retardates on Day 2 as a function of social class.

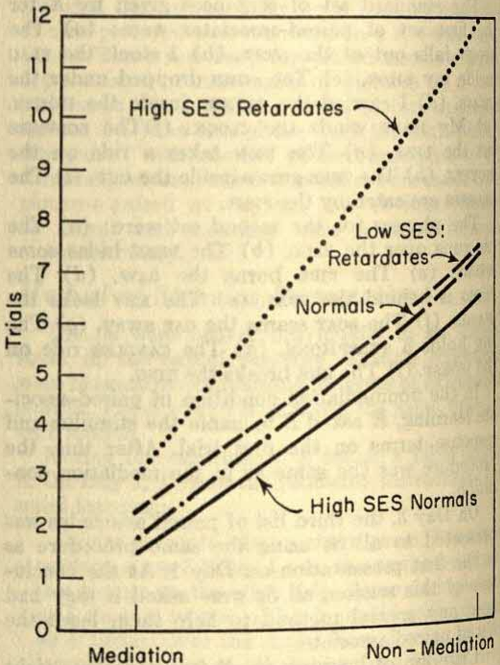


FIG. 2. Trials to criterion of four groups for mediation and nonmediation on Day 2.

Table 2 also shows the means and standard deviations for all groups on Day 3. The main treatment effect is not significant. There is no significant difference ($F = 0.68$, $df = 1/71$, $p = >.10$) in the learning of paired associates on Day 3 between groups who had received mediation instruction on Day 2 as compared to the groups who had not received instructions. Main effects were significant for IQ ($F = 7.92$, $df = 1/71$, $p < .01$) and SES ($F = 7.90$, $df = 1/71$, $p < .01$). The normal group required significantly fewer trials to criterion than the retarded. The low-SES group surpassed the high-SES group in

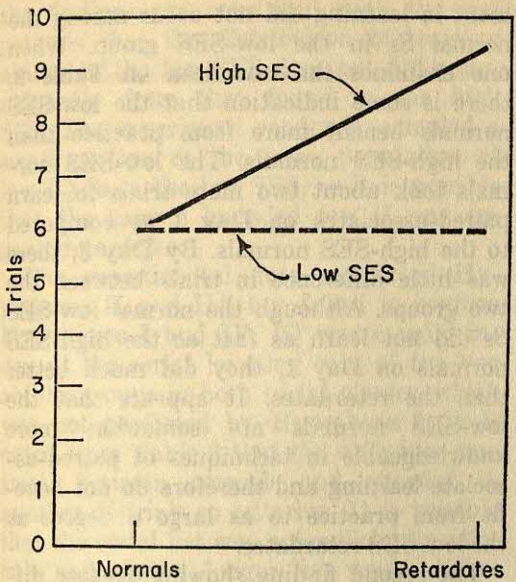


FIG. 3. Trials to criterion of high and low SES groups on Day 3 as a function of IQ.

speed of learning. There was a significant SES × IQ interaction ($F = 8.19$, $df = 1/71$, $p < .01$) shown in Figure 3. Pairwise comparison by the Scheffé technique revealed that the locus of the SES effect was between the two retarded groups. Low-SES retarded *S*s required fewer trials to criterion than high-SES retarded *S*s. Again, the latter group did not differ from their normal counterpart in performance. This was in contrast to the normal and retarded *S*s from high SES who continued to show a significant difference in performance. The high-SES retardates took about twice the number of trials to reach criterion as the other groups.

Intercorrelations among Variables

Product moment correlations were determined for MA, IQ, and learning tasks. Due to the use of extreme groups, correlations are probably spuriously high but they reflect the pattern of correlations that would be found in the total population. Most interesting are the intercorrelations among measures in the high- and low-SES groups reported in Table 3. IQ correlated significantly with all learning tasks in the high SES but not in the low SES. With the exception of serial learning, the differences between the correlation coefficients found in the two SES groups are significant ($p < .05$). Therefore, one can conclude that there is a positive relationship between IQ and paired-associate learning for high-SES *S*s but not for low-SES *S*s.

DISCUSSION

Two important findings resulted from the experiment. One, there is a significant

TABLE 3
INTERCORRELATIONS AMONG VARIABLES FOR HIGH- AND LOW-SES GROUPS
(Above diagonal High SES; below diagonal Low SES)

	1	2	3	4	5	6	7
	MA[a]	MA[b]	IQ	Serial learning[c]	Paired associates (Day 1)[c]	Paired associates (Day 2)[c]	Paired associates (Day 3)[c]
1		.83**	.79**	.44**	.31	.39**	.54**
2	.86**		.70**	.32*	.32*	.41**	.64**
3	.82**	.68**		.43**	.52**	.41**	.60**
4	.19	.20	.22		.58**	.19	.45**
5	.22	.18	.16	.03		.27	.57**
6	.10	.00	.01	.00	.14		.42**
7	.14	.13	.08	.06	.53**	.29	

[a] MA based on Stanford Binet Intelligence Scale and Kuhlman-Anderson Intelligence Test.
[b] MA based on Peabody Picture Vocabulary Test.
[c] Since variables 4, 5, 6, 7 are trials-to-criterion scores, these variables have been reflected to yield positive correlations between the psychometric tests and the learning measures.
* $p < .05$.
** $p < .01$.

difference in the learning ability of retardates from high SES as compared to retardates from low SES. Two, IQ is a better predictor of learning ability in the high-SES groups than in the low-SES groups.

The retarded groups did not differ in performance on the first day of the experiment. By the end of the week, the two groups were strikingly different in rate of learning. The low-SES retardate was learning paired-associates at a much faster rate than he did earlier in the experiment; the high-SES retardate was continuing to learn at the same slow rate. However, the surprising aspect of this result is that the improvement in learning ability among the low-SES retardates cannot be attributed to instructions in mediation as the control group showed as much improvement as the experimental group on Day 3.

What caused the sudden improvement among the low-SES retardates? These retardates were not able to indicate any particular method they used to learn the lists when questioned at the conclusion of the experiment. Common replies were, "I said them over and over," "I memorized them," "I don't know." The first hypothesis that comes to mind is that the improvement is due to the effects of practice and learning to learn.

All four groups showed some improvement due to practice, but why should the effect be greatest among the low-SES retarded? Past research (Covington, 1962; Haggard, 1954) has shown that lower-class *S*s benefit most from the opportunity to practice or become familiar with a task. This is generally explained by the lack of opportunity among lower-class *S*s to become familiar with different ways of responding to a variety of stimuli so their performance will be markedly inferior to upper-class *S*s in many learning situations. Once the lower-class *S*s have the opportunity to practice, they will benefit more than the upper-class *S*s. The upper-class *S* benefits less from practice as his performance is nearer the limits of his learning ability during the initial trials. However, this does not explain why a similar improvement in learning did not occur among the normal *S*s in the low-SES group. When one examines the raw data on Table 2, there is some indication that the low-SES normals benefit more from practice than the high-SES normals. The low-SES normals took about two more trials to learn paired-associates on Day 1 as compared to the high-SES normals. By Day 3, there was little difference in trials between the two groups. Although the normal low-SES *S*s did not learn as fast as the high-SES normals on Day 1, they did much better than the retardates. It appears that the low-SES normals are somewhat more knowledgeable in techniques of paired-associate learning and therefore do not benefit from practice to as large a degree as the low-SES retardates.

The second finding showing greater difference in learning ability between IQ groups in the high-SES than in the low-SES sample is contrary to usual reports. On Day 1 tasks, normal IQ *S*s learned faster than retardates in both SES groups. Over the rest of the tasks, there continued to be IQ differences in learning ability among high-SES *S*s, but not among low-SES *S*s where differences in learning ability gradually disappeared. Why should IQ be a better predictor of learning ability in the high-SES than in the low-SES group?

As indicated earlier, education and occupation are crude measures of SES. Probably, there is much more variability in the kinds of environmental experiences in the low SES than in the high SES. The work of some investigators (Deutsch, 1963; Eells, 1951) suggests that children from upper-class homes have a more common environment due to prolonged schooling, more family stability, and more exposure to books, magazines, and newspapers. The environment of lower-class children is less uniform due to inconsistencies in schooling, more mobility, less family stability, and less exposure to books and magazines. Thus, the performance of the low-SES *S*s will be more variable and unpredictable.

These results are not meant to imply that there are not any organic retardates among the low-SES population. Rather

the findings are consistent with the hypothesis that the distribution of learning ability in the low SES does not differ significantly from that found in the high SES. The experiment randomly sampled a small portion of the large population of low-IQ *S*s in the low SES, so it is possible that the sample included few if any organic retardates.

The experimental data supported Hypotheses II and III, but failed to confirm Hypotheses I and IV. IQ was a valid index of the serial learning rate of the two ability groups in both social classes, which was not predicted by Hypothesis I. It is noted from Table 2 that the difference was small between normals and retarded in number of trials to criterion. It is possible that the serial list was not sufficiently difficult and that a longer list would differentiate more clearly among groups. An alternate hypothesis is that serial learning may also depend on certain subskills which have not been analyzed as yet. Thus, the low-SES retardate may have been just as handicapped by environmental deficiencies in certain subskills required by serial learning as he was on the paired-associate task. Data on the learning of several serial lists might shed some light on this area.

Paired-associates closely reflected IQ differences in both the high- and low-SES groups as predicted by Hypothesis II. The effect of mediation instruction was to greatly facilitate paired-associate learning, confirming Hypothesis III. However, the use of verbal links did not wipe out IQ differences as normals continued to maintain their superiority.

Data failed to support Hypothesis IV as there was no evidence of the mediation set being transferred to the learning of a new list of paired-associates on Day 3. However, one cannot conclude that retardates are basically deficient in the ability to transfer skills acquired in one situation to another as normal *S*s did not show any superiority in retaining the mediation habit. The lack of difference between mediation and nonmediation groups on Day 3 can be explained in several ways. First, one training session is probably not sufficient to inculcate the habit of forming verbal chains between corresponding pairs on a paired-associate task. All of the retarded and many of the normals were unable to describe any method which they had used to learn the paired-associates on Day 3. Second, some of the control *S*s spontaneously developed habits of making verbal links which may have canceled any special benefits which had occurred in the experimental group. For example, among the high-SES normals half the *S*s in the mediation group reported mediating, but just about the same number of the nonmediation group described the use of some kind of mnemonic device. Third, instruction in the use of verbal mediators may have interfered with previously established habits of learning. For example, one *S* who reported attempting to make up sentences to link the pairs, as he had been taught on Day 2, took more trials to learn the paired-associate list on Day 3 than he had on Day 1. Fourth, the mediation group may not have been able to make up as elaborate verbal chains as were provided by *E* on Day 2. Recent data by Rohwer (1965) indicate that the structure of language plays an important role in the amount of facilitation produced by verbal chains. In their role as connectives, verbs produced the most facilitation and conjunctions the least, while prepositions were somewhere in the middle. Sentences provided by *E* in the present investigation were made up of from five to seven words and always included a verb.

The present experiment has shown that there are differences in learning abilities among mentally retarded which need to be considered in planning their educational program. Low measured IQ and a history of educational failure should not be the only criteria for placement in a special class. An additional measure should consist of verifying whether in fact the mentally retarded cannot learn from appropriate experiences. The customary assessment fails to take into account whether these persons are low in IQ and achieve-

ment due to organic deficiencies or due to an environment which has failed to provide them with the necessary knowledge and skills.

REFERENCES

COVINGTON, M. Some effects of stimulus familiarization on discrimination performance. Unpublished doctoral dissertation. University of California, 1951.

CRONBACH, L. J. *Essentials of psychological testing.* New York: Harper, 1960.

DEUTSCH, M. The disadvantaged child and the learning process. In A. H. Passow (Ed.), *Education in depressed areas.* New York: Teachers College, Columbia University, 1963.

DUNN, L. M. *Peabody picture vocabulary test.* Minn: American Guidance Service, 1965.

EELLS, K., DAVID, A., HAVIGHURST, R. J., JERRICK, V. E., & TYLER, R. *Intelligence and cultural differences.* Chicago: University of Chicago Press, 1951.

GRIFFITH, B. C., SPITZ, H. H., & LIPMAN, R. S. Verbal mediation and concept formation in retarded and normal subjects. *Journal of Experimental Psychology,* 1959, **58,** 247–251.

HAGGARD, E. A. Social status and intelligence. *Genetic Psychology Monographs,* 1954, **49,** 141–186.

HAYS, W. L. *Statistics for psychologists.* New York: Holt, Rinehart & Winston, 1963.

JENSEN, A. R. Rote learning in retarded adults and normal children. *American Journal of Mental Deficiency,* 1965, **69,** 328–334.

JENSEN, A. R. Social class and verbal learning. In M. Deutsch, A. R. Jensen, & I. Katz (Eds.), *Social class, race and psychological development.* New York: Holt, Rinehart, & Winston, 1967.

JENSEN, A. R., & ROHWER, W. D., JR. Verbal mediation in paired-associate and serial learning. *Journal of Verbal Learning and Verbal Behavior,* 1963, **1,** 346–352. (a)

JENSEN, A. R., & ROHWER, W. D., JR. The effect of verbal mediation on the learning and retention of paired-associates by retarded adults. *American Journal of Mental Deficiency,* 1963, **68,** 80–84. (b).

JENSEN, A. R., & ROHWER, W. C., JR. Syntactical mediation of serial and paired-associate learning as a function of age. *Child Development,* 1965, **36,** 601–604.

MCCARTHY, D. Language development in children. In L. Carmichael (Ed.), *Manual of child psychology.* New York: Wiley, 1954. Pp. 492–631.

RIEBER, M. Verbal mediation in normal and retarded children. *American Journal of Mental Deficiency,* 1964, **68,** 634–641.

ROHWER, W. D., JR. The verbal facilitation of paired-associate learning. Unpublished doctoral dissertation, University of California, 1965.

SARASON, S. B., & GLADWIN, T. Psychological and cultural problems in mental subnormality: A review of research. *Genetic Psychology Monographs,* 1958, 57.

SIEGEL, S. *Nonparametric statistics.* New York: McGraw-Hill, 1956.

(Received March 27, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 2, 111–118

DEVELOPMENTAL STUDY OF THE MEANING OF ADVERBIAL MODIFIERS[1]

W. L. BASHAW

AND

HARRY E. ANDERSON, Jr.[2]

University of Georgia

The development of the meaning of adverbial modifiers was studied by pair-comparison and ranking methods. The adverbs studied consisted of slightly, somewhat, rather, pretty, unmodified form, quite, decidedly, unusually, very, and extremely. The scaling tasks were administered to Ss in grades 1, 2, 3, 4, 5, 6, 8, 10, 12, and college. Obtained scale values were highly reliable. Scalability was seen to relate positively to age-grade classification. Primary-grade Ss identified at least 3 adverb groups while adults identified about 6 groups. Correlations of scale values of primary-grade Ss with college Ss ranged from .74 to .94. All other groups yielded correlations with college data above .90. Some words were seen to shift in meaning as a function of age-grade group. Results are interpreted in terms of applications to general scaling methodology, measurement methodology with young children, and research in language development.

The study of the quantification of the meaning of words has become increasingly important since Osgood (1952) introduced the semantic differential technique for studying connotative meaning. However, some word sets, such as certain adverbs and adjectives, apparently can be ordered along specific continua and might have general quantifiable denotative meaning.

Reports of the actual scaling of word meaning are rare. Darley, Sherman, and Siegel (1959) scaled the abstraction level of nouns, adjectives, and verbs with a high degree of reliability and Cliff (1959) studied extensively the scaling of adverbial modifiers. Cliff was interested primarily in a theory of the use of adverb scale values as multipliers of the scale values of adjectives. This latter study is of major interest.

Cliff (1959) studied a set of 10 adverbial modifiers in combination with 15 adjectives. He administered the adverb-adjective combinations as a successive-intervals judging task using an 11-point response continuum. His subjects (*S*s) were college elementary psychology students. The adverbs generally scaled in the following order (from low to high intensity of modification): (*a*) slightly, (*b*) somewhat, (*c*) rather, (*d*) pretty, (*e*) unmodified form, (*f*) quite, (*g*) decidedly, (*h*) unusually, (*i*) very, and (*j*) extremely. Cliff's average scale values (reported by Dudek, 1959) were, respectively, .55, .69, .86, .92, 1.00, 1.07, 1.20, 1.30, 1.30, and 1.53. (The scale is transformed to make the unmodified form the unit of the scale.) Cliff found that, usually, the choice of adjective did not affect the ordering of the adverbs. One notable exception was the adverb "pretty." This adverb was seen to make unfavorable adjectives more extreme and to make favorable adjectives less extreme.

A third relevant paper (Dudek, 1959) was methodological in nature. Dudek compared scale values of the adverbs studied by Cliff as determined by successive intervals and by the constant-sum method. The two methods did not yield exactly

[1] The research and development reported herein was performed pursuant to a contract with the United States Department of Health, Education and Welfare, Office of Education, under the provisions of the Cooperative Research Program. This research was conducted by the Evaluation Division of the Research and Development Center in Educational Stimulation.

[2] The investigators wish to express their appreciation to the following persons for their kind assistance: Jasper Griffin, Superintendent, Cobb County Schools, Gaynell Walker, Coordinator of Elementary Education, J. Milton Lewis, Coordinator of Secondary Education, and to the many participating principals and teachers from the Cobb County School System.

equivalent results; however, results were similar. Dudek's average scale values for Cliff's adverbs were, respectively, .74, .83, .94, 1.02, 1.00, 1.28, 1.47, 1.68, 1.66, and 2.07. Note that in only two cases are these values not in the same order as Cliff's values. The two reversals are pairs of adverbs that are extremely close in meaning as evidenced by the similarity in magnitude of the scale values.

Allison (1963) showed the importance of the study of adverb meaning by applying Cliff's results to semantic differential methodology. Allison attempted to increase the range of scale values by identifying scale end points with adverbs of high modification.

The present investigation extends the study of adverb meaning downward to the first-grade level. It was hypothesized that the ability to discriminate between adverbial modifiers increases with chronological age and educational level. For example, it was believed that children in the primary grades could provide at least polar classification of adverbs—that is, they could identify adverbs as extreme modifiers or moderate modifiers. Older children were expected to make finer distinctions. Dudek's (1959) report indicates that college students distinguish about eight points (out of a possible 10) on an adverbial intensity of modification scale; however, this conclusion was not based on statistical tests of the differences in preferences for members of pairs. It was predicted that a monotonically increasing function will be observed relating age-grade group to the ability to discern adverbial modifiers. It was also predicted that a monotonically decreasing function will be observed relating age-grade group to the dispersion values of the adverbs.

An important outcome is the determination of differences in semantic meaning that various age groups give to the several adverbs. Another important outcome is insight into appropriate methodology for presenting abstract scaling tasks to primary grade children.

Method

Characteristics of the Population

The population of interest is the general public-school membership. It was desired to choose an upper-level group that consisted of highly educated persons rather than members of the general public since this group was used for the "criterion ranking." For this reason, college sophomores were considered to be a good choice. Data obtained on college sophomores could also be compared to that of Cliff and Dudek to determine if generalization to other groups and methodologies can be made safely at the upper range of ability and chronological age. If these comparisons are favorable, then one can have more faith in the generalizations drawn from data from younger groups.

The population was restricted to students who were within the typical chronological age range for their grade group. Thus, slow learners and students who have skipped grades were deleted. This step was taken to assure that grade groups will not have extreme variability with respect to chronological age. Thus, grade groups can be treated roughly as age groups.

Characteristics of the Sample

Intact classroom groups were used for the entire study. Sampling of classroom groups was incidental within the limitations specified below. About 180 *S*s were chosen from each of the grade groups 1, 2, 3, 4, 5, 6, 8, 10, 12, and 13. No limitation was imposed on the membership of the sample other than each child must have had a chronological age within 6 months of the modal age for his grade group. This limitation did not apply to the college sample. In some grades, the limitation resulted in the discarding of considerable data. Final sample sizes ranged from 116 to 184.

Care was taken to choose schools which were not expected to be atypical in any way related to the language achievement of the students. Classroom groups chosen were not grouped on any scheme of homogeneous grouping. Public schools were chosen such that the elementary schools were feeder schools to the junior high schools and the junior high schools were feeder schools to the senior high school. Hopefully, this selection resulted in some control over socioeconomic differences among the grade groups. All schools were located in a suburban area near Atlanta, Georgia.

General Design

The dependent variables under study are estimates of the various adverb-scale parameters—namely, scale values, scale dispersions, and the number of discernible scale points for the 10 adverbs used by Cliff and Dudek. The independent variable is age-grade group.

The 10 adverbs were used with the adjective "large" to form the adverbial phrases used in the

scaling tasks. The neutral adjective "large" was chosen so that nonabstract examples could be used by the examiner in the task administration and by *S*s in their consideration of the various stimuli pairs. Each adverb-pair was presented in both orders of presentation for most *S*s (exceptions are noted below). Thus, most *S*s responded to a randomized list of 90 pairs, or two complete replications. Also, each *S* was presented a randomized list of 10 adverbial phrases and was asked to rank these from low to high. The pair-comparisons always preceded the ranking task.

Materials and methods were modified as necessary for the younger *S*s. Primary-grade *S*s received materials printed with large type and instructions were simplified for them. The examiner read aloud any words upon request of middle-grade children. All tasks were read aloud to all *S*s in the first and second grades. Example tasks were used to verify that all children understood the nature of the pair-comparison tasks and understood the concept of "largeness." During the second-grade administration it was observed that the *S*s took an unusually large amount of time to complete all tasks. Therefore, part of the second-grade group and all of the first-grade group were asked to complete only one replication of the pair-comparisons tasks.

Analyses

Scale values. Scale values of each adverb were determined for each age group and for each replication by pair-comparison methodology (Edwards, 1957, pp. 19–36).

Dispersion values. Dispersion values for each scale value were determined for each age group by pair-comparison methodology (Edwards, 1957, pp. 58–66).

The number of discernible scale points. This value was determined for each age group based on significance tests of differences in rank-totals. This procedure is based on the relationship of pair-comparison methodology to analysis-of-variance of ranked data (Dunn-Rankin, 1965; Dunn-Rankin & Wilcoxon, 1966).

Scalability Index. It was desired to have an index that would show the degree to which the adverbs were scaled by each group of *S*s. The coefficient of agreement (u; Edwards, 1957, pp. 76–78) is one such index and it is reported. However, a particular hypothesis that was to be tested in this study dealt with the number of distinguishable scale points in the adverb-scale of each group of *S*s and this hypothesis requires a different statistic.

The initial analysis determined the adverb pairs that were statistically distinguishable using Dunn–Rankin's (1965) technique. This analysis, in some cases, divided the 10 adverbs into mutually exclusive sets. Usually, however, sets were not mutually exclusive, but overlapped. For example, "slightly" (sl) and "somewhat" (so) might be perceived as synonomous, "somewhat" and "rather" (r) might be perceived as synonomous, but "slightly" and "rather" might be perceived as different. The Dunn–Rankin multiple-range test would show sl = so, so = r, and sl < r. This result should be classified as "more scalable" than the result sl = so = r, but it should be classified "less scalable" than the result sl < so < r.

It was decided to use a simple coefficient of scalability that reflects the number of discernible pairs. The coefficient decided upon was the ratio of discernible adverb pairs to the total number of pairs (which was 45 in all cases). This ratio was expressed as a percentage and was called SI for "Scalability Index."

The coefficient SI has a possible range of 100. SI has a minimum of zero if no pairs yield statistically significant differences and SI takes the maximum of 100 if all pairs are mutually significantly different. Only in the latter case does one get a total ordering of the stimuli.

The investigators are not proposing SI as a new statistic to replace coefficients like the coefficient of agreement or Guttman's reproducibility coefficient. It is a statistic that, in the judgment of the investigators, adequately summarizes the results of each group of *S*s so that a test of the developmental hypothesis of adverb meaning can be accomplished.

RESULTS

Results are presented for each specific objective of the study. The first two objectives were to determine, by pair-comparison technique, the scale values and discriminal dispersions of each adverb for each age-grade group. The scale values for each analysis appear in Table 1 and the corresponding discriminal dispersions appear in Table 2. The adverbs are in the same order as listed by Cliff (1956) and Dudek (1959). The abbreviations are slightly (sl), somewhat (so), rather (r), pretty (p), unmodified form (l), quite (q), decidedly (d), unusually (u), very (v), and extremely (e). The correlations between scale values of replications within grade groups provide estimates of all reliabilities of the pair-comparison data. The lowest of the 11 reliabilities was .92. Seven of the 11 correlations were .97 or higher. The correlations indicate a high reliability for scale values, even for the primary grades. No check was made of the correlations between rankings for individuals.

The third and fourth objectives were to determine the number of discernible scale points at each grade level and to determine at what age children can first begin to

TABLE 1
PAIR-COMPARISON SCALE VALUES FOR 10 ADVERBS BY GRADE LEVEL

Grade	N	Adverb									
		sl	so	r	p	l	q	d	u	v	e
1	187	−.23	−.26	−.08	−.20	−.34	−.36	.13	.23	.62	.48
2a	102	−.22	−.46	−.10	−.34	−.32	−.35	.20	.12	.63	.84
2b	82	−.48	−.26	−.23	−.26	−.29	−.11	.29	.16	.57	.62
2c	82	−.26	−.26	−.24	−.40	−.33	−.18	.24	.15	.56	.71
3a	158	−.69	−.36	−.33	−.35	−.24	−.03	.21	.24	.65	.90
3b	158	−.56	−.42	−.31	−.41	−.21	−.12	.20	.18	.60	1.05
4a	143	−.93	−.75	−.28	−.33	.04	.07	.14	.29	.75	1.00
4b	143	−.84	−.63	−.33	−.36	−.02	.03	.10	.22	.83	.99
5a	116	−1.19	−.76	−.26	−.46	−.23	.13	.04	.51	.79	1.44
5b	116	−.99	−.70	−.34	−.42	−.35	.16	−.08	.48	.82	1.42
6a	128	−.97	−.53	−.11	−.23	−.37	.24	.02	.58	.57	.79
6b	128	−.85	−.79	−.37	−.40	−.41	.07	−.25	.49	1.02	1.49
8a	162	−1.07	−.51	−.05	−.26	−.47	.25	−.02	.81	.52	.81
8b	162	−.81	−.71	−.38	−.46	−.61	.14	−.38	.79	.96	1.46
10a	123	−1.25	−.72	−.15	−.22	−.43	.23	−.06	.89	.66	1.04
10b	123	−1.04	−.84	−.46	−.53	−.56	−.01	−.28	1.04	1.04	1.65
12a	164	−1.48	−.92	−.12	−.37	−.34	.26	.07	1.00	.70	1.21
12b	164	−1.10	−1.08	−.45	−.55	−.49	.07	−.20	1.01	1.01	1.77
13a	163	−1.40	−1.01	−.49	−.51	−.40	.35	.30	1.15	.68	1.34
13b	163	−1.14	−1.00	−.55	−.53	−.47	.17	.21	1.08	.83	1.40

discriminate between the adverbial modifiers. The numbers of discernible scale points were determined from the ranking data, and, in the case of Grades 1 and 2 only, from the pair-comparison data. Tables 3 and 4 show the relevant results. Table 3 shows the rank-order of all adverbs for each group based on the ranking data.

TABLE 2
PAIR-COMPARISON DISCRIMINAL DISPERSION VALUES FOR 10 ADVERBS BY GRADE LEVEL

Grade	N	Adverb									
		sl	so	r	p	l	q	d	u	v	e
1	187	.83	.99	.66	1.37	1.06	1.14	1.00	.83	1.19	.91
2a	102	1.07	1.18	.73	1.19	.90	1.20	.98	.82	.92	1.01
2b	82	1.14	.76	1.50	.84	.75	1.03	1.35	.86	.58	1.20
2c	82	1.95	.70	.75	.80	.78	1.41	1.13	.90	.48	1.11
3a	158	1.72	1.07	.94	.74	.73	.55	.92	1.14	.53	1.67
3b	158	2.45	.73	.91	.71	.63	.70	1.02	1.25	.53	1.07
4a	143	1.50	.89	.70	.56	.97	.67	1.37	1.39	.51	1.43
4b	143	1.83	1.16	.55	.58	.50	.61	1.39	1.42	.46	1.51
5a	116	1.62	1.07	.69	.61	.90	.61	1.12	1.41	.54	1.43
5b	116	2.07	1.18	.60	.65	.73	.64	.91	1.19	.48	1.54
6a	128	1.40	.88	1.01	.69	.68	1.41	1.10	1.57	.49	.77
6b	128	.89	.85	1.12	1.24	1.17	.53	1.32	.98	.39	1.51
8a	162	1.57	.86	.93	.70	.77	.93	1.09	1.63	.42	1.10
8b	162	1.02	1.17	.94	.91	1.21	.50	1.12	.77	.45	1.91
10a	123	1.62	.98	.69	.73	.79	.79	1.00	1.51	.41	1.48
10b	123	1.41	1.48	.81	.63	1.02	.35	1.12	.94	.49	1.74
12a	164	1.59	.86	.74	.68	.75	.73	1.17	1.47	.52	1.50
12b	164	1.53	1.32	.85	1.06	.72	.43	.97	.92	.39	1.81
13a	163	2.18	1.08	.55	.58	.53	.56	.73	1.55	.51	1.71
13b	163	2.13	1.40	.68	.69	.56	.36	.52	1.13	.52	2.01

TABLE 3

SUMMARY OF ADVERB-SCALING RESULTS—FINAL ORDERS AND INDISTINGUISHABLE ADVERBS

Group	Ranking data results									
1	R	L	Q	Sl	P	So	D	U	V	E
2	R	Q	L	Sl	P	So	U	D	V	E
3	Sl	L	R	P	So	Q	V	D	U	E
4	Sl	L	So	R	P	Q	D	U	V	E
5	Sl	L	So	R	P	D	Q	V	U	E
6	Sl	L	So	R	P	D	Q	V	U	E
8	Sl	L	So	P	R	D	Q	V	U	E
10	Sl	So	L	P	R	D	Q	V	U	E
12	Sl	So	P	L	R	D	Q	V	U	E
13	Sl	So	R	L	P	Q	D	V	U	E
	Pair-comparisons results									
1	P	R	Q	L	So	Sl	D	U	E	V
2	So	P	L	Q	Sl	R	U	D	V	E

Note.—Indistinguishable pairs are underlined—solid lines show 95% confidence level and dotted lines show 90% confidence level.

Indistinguishable pairs are underlined. The numbers of distinguishable pairs, the corresponding SI values, and the coefficients of agreement, u, appear in Table 4.

It is apparent from these data that a large number of primary grade students can properly classify the adverbs into at least three sets. The number is sufficiently large in the first grade to yield 26 significant differences in scale values at the first-grade level. It is apparent, from the low u-values, that there is also considerable *disagreement* as well as agreement in the primary years. The differences between students at this level who can and who cannot correctly perform the scaling tasks should become the subject of a series of highly interesting and informative studies. The fifth objective was to determine the relationship of age-grade group to the number of discernible scale points.

The regression of adverb distinguishability on age-grade group appears to yield basically a monotonically increasing function. Only two pairs of SI-values are not in monotonic order and only one u-value is out-of-place in regard to a monotonic regression function.

The pair-comparison data of the two youngest groups were also analyzed by Dunn-Rankin's technique and these results are also in Table 3. The results were slightly better than the ranking results.

TABLE 4

SCALABILITY INDEXES BY AGE-GRADE GROUPINGS

Group	N	Ranking Data		Pair-comparison data—u[b]
		Distinguishable pairs	SI(%)[a]	
1	187	26	58	.14
2	184	27	60	.14
3	158	33	73	.24
4	143	35	78	.32
5	116	37	82	.42
6	128	37	82	.33
8	162	39	87	.38
10	123	38	84	.46
12	164	40	89	.51
13	163	39	87	.59

[a] Scalability Index—ratio of the number of distinguishable pairs to the possible maximum of distinguishable pairs (45) expressed as a percentage.

[b] Coefficients of agreement are based on scale values averaged over replications.

TABLE 5
Correlations of Pair-Comparison Scale Values for Each Group with Adult Scale Values

Group	13a	13b
1	74	81
2a	76	83
2b	87	91
2c	81	86
3a	92	94
3b	88	92
4a	93	94
4b	91	93
5a	96	96
5b	94	96
6a	97	96
6b	91	94
8a	97	95
8b	91	95
10a	97	97
10b	94	97
12a	98	97
12b	95	98

Note.—An r of .63 is significant at the .05 level of confidence for $df = 8$.

The number of distinguishable pairs was 29 for the first grade (SI = 64%) and 28 for the second grade (SI = 62%).

The sixth objective deals with the relationship of dispersion values of adverbs to age-grade group. It was hypothesized that the dispersion values would tend to increase with age. However, an inspection of Table 2 reveals no evidence of trends. One could conclude that the scale values are just as variable for adults as for primary grade children. However, the u-values of Table 4 reveal the degree to which adults are more in mutual agreement than were the younger children.

The final objective was to compare the scale values of each grade to the scale values of adults (Group 13). The scale values of each group were correlated with the two sets of scale values for the adults. These correlations appear in Table 5.

The relationship of the present results to the results presented by Cliff (1959) and Dudek (1959) are of considerable interest because of the differences in populations sampled and in scaling methodology. The results of Cliff and Dudek were based on data from college sophomores, so the intercorrelations of their scale values and that of the sophomores in this study give evidence of generalizability across samples and methodologies. Correlations were calculated between the scale values for the two replications of Group 13 and the average scale value of Cliff's "good" adjectives, Cliff's "bad" adjectives, Cliff's overall averages, Dudek's "favorable" adjectives, Dudek's "unfavorable" adjectives, and Dudek's overall averages. Each of these 12 correlations was .96 or higher, indicating extremely high similarity among the results of the present study and the two previously reported studies. The 6 × 6 intercorrelation matrix of Cliff's and Dudek's results (unreported) had no value below .94.

The correlations of the results of each grade group with the Cliff and Dudek results are also of interest as evidence of generalizability of the developmental data. Since the adult intercorrelations were so high, the correlations of each group with the Cliff and Dudek results were quite similar to the corresponding correlations with Group 13 so they are not reported. However, the correlations did give some additional information. There was a tendency for the values to be slightly larger for "good" and "favorable" adjectives than for "bad" or "unfavorable" adjectives. The present task involved a neutral adjective "large" which is apparently perceived as more like a positive than a negative adjective.

Discussion

The results of the study were surprising in one regard—the younger Ss performed the task with considerably more skill than was anticipated. The reliabilities and the correlations with adult values were much higher than was expected. On the other hand, the adult data, although highly reliable and highly consistent with results reported in other studies, were not as internally consistent as was expected. This claim is based on the low number of zero

and near-zero frequencies in the preference matrix for adults. In any case, all groups yielded highly reliable scale values.

The investigators were especially pleasantly surprised to see the ease with which the younger children handled the complex ranking task. The *S*s did have considerable practice at making pair-wise decisions; however, in spite of this the full ranking task was expected to be extremely difficult. There is some weak evidence that the pair-comparison task did yield better data. The SI values were 58% (first grade) and 60% (second grade) for total ranking as compared to 64% and 62%, respectively, for pair-comparisons.

One major value of the study is the determination of relative meaning of the adverbs as a function of age and training. The relative meanings can best be seen in Table 3, which lists the adverbs in increasing order of scale value. Note the shift in meaning of the words "somewhat" (so) and "slightly" (sl) from neutral at the primary level to extreme at the adult level.

The word "very" is also of high interest due to the large use of it in defining responses in attitude scaling. "Very" is seen as less strong than either "extremely" or "unusually" for grades higher than five. "Very" is seen as less strong than "extremely" above Grade 2. But Grades 1 and 2 equate "extremely" and "very."

It is expected that many scaling studies can be improved by using the reported scale values to increase the variance of responses and clarify anchoring definitions. The researcher would need to choose adverb scale values corresponding to the age and training of his *S*s. Note that if one wanted to scale objects according to "size," one could list responses from "extremely small" through "slightly small" and "slightly large" up to "extremely large." Cliff (1959) indicated considerable invariance of adverb scale values with regard to the adjective modified, so one would expect the values reported herein to apply to adjectives other than "large." The use of the various adverb scale values to space responses conceivably could help assure that response intervals were indeed on a meaningful interval. In some studies, one could possibly apply the multiplication rules of Cliff (1959) for determining adverb-adjective-pair scale values for the same purpose. The scale values can also be used creatively to break down lack of variation due to generosity effect. For example, an employee could be rated from "slightly good" to "very good" instead of from "very bad" to "very good."

In general, one can conclude that there is need for considerable care in the construction and administration of scaling tasks for young children. This conclusion is based largely on the dearth of methodological work at this age level. It is based in part on the findings of some differences in the meanings that various adverbs have for different age groups and for individuals within age groups. The meaning of words that are not routinely taught in vocabulary instruction cannot be assumed to have identical meaning to all persons, although the adverbs studied do have fairly stable meaning across age-grade group.

There is also need for the study of other word-types and how the usage and meaning of words change with age and training. Perhaps the study of abstraction levels of words (Darley, Sherman, & Siegel, 1959) can provide a model for this needed body of research.

REFERENCES

ALLISON, R. B. Using adverbs as multipliers in semantic differentials. *Journal of Psychology,* 1963, **56,** 115–117.

CLIFF, N. Adverbs as multipliers. *Psychological Review,* 1959, **66,** 26–44.

DARLEY, F. L., SHERMAN, D., & SIEGEL, G. M. Scaling of abstraction level of single words. *Journal of Speech and Hearing Research,* 1959, **2,** 161–167.

DUDEK, F. J. A comparison of scale values for adverbs determined by the constant-sum method and a successive intervals procedure. *Educational and Psychological Measurement,* 1959, **19,** 539–548.

DUNN-RANKIN, P. The true probability distribution of the range of rank totals and its applica-

tion to psychological scaling. Unpublished doctoral dissertation, Florida State University, 1965.

Dunn-Rankin, P., & Wilcoxon, F. The true distributions of the range of rank totals in the two-way classification. *Psychometrika*, 1966, **31**, 573–580.

Edwards, A. L. *Techniques of attitude scale construction.* New York: Appleton-Century-Crofts, 1957.

Osgood, C. E. The nature and measurement of meaning. *Psychological Bulletin*, 1952, **49**, 197–237.

(Received March 30, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 2, 119–127

PREDICTING PERFORMANCE IN STUDENT TEACHING FROM THE CALIFORNIA PSYCHOLOGICAL INVENTORY

HARRISON G. GOUGH
University of California, Berkeley

GLENN W. DURFLINGER
University of California, Santa Barbara

AND ROBERT E. HILL, JR.
Ball State University

An equation to predict teaching effectiveness from the California Psychological Inventory (CPI) was developed on a sample of 293 students (215 females, 78 males). The equation included scales for socialization, good impression, achievement via conformance, psychological-mindedness, and flexibility, combined in this form: 14.743 + .334So − .670Gi + .997Ac + .909Py − .446Fx. Cross-validation of the equation on 17 higher-rated vs. 17 lower-rated student teachers gave a t ratio of 4.27 ($p < .001$), and a biserial correlation of +.44. Conceptual analysis of the dimension defined by the equation suggested personological bases of conscientiousness vs. undercontrol of impulse for males, and of resoluteness versus indifference for females.

The prediction of performance in student and/or professional teaching is one of the long-standing, unsolved, and perhaps (some would say) unsolvable problems of educational psychology. Still, in spite of its difficulty (cf. Barr, 1958), the issue continues to excite the interest of researchers, and significant publications on teaching effectiveness (cf. Biddle & Ellena, 1964; Ellena, Stevenson, & Webb, 1961; Ryans, 1960) continue to attend to it. Gage's *Handbook* (Gage, 1953), in fact, carries two chapters in which important discussions of the problem are to be found (Getzels & Jackson, 1963; Stern, 1963). Thus, although a certain weariness is understandable, there is no need as yet to abandon inquiry.

Specification of a criterion of effectiveness has been a principal difficulty, and in the judgment of Getzels and Jackson perhaps the major stumbling block. Traditionally, there have been three ways of establishing the criterion: first, from an evaluation of the scholastic achievement of students; second, from ratings or judgments by supervisors who have observed the teacher in the classroom; and third, from ratings or judgments furnished by the students themselves.

There are obvious problems in using the first method (e.g., controlling for intellectual level and past experience of the students), and current opinion (cf. Chauncey & Dobbin, 1964) is generally against its employment.

The second method is more favored, particularly if observations can be systematized so that both intra- and interjudge comparisons can be made. Various rating forms which fulfill these requirements have been developed. For example, Durflinger (1963) has devised a 41-item teacher-evaluation scale, and Michaelis (1954) has described a progress report form capable of yielding quantitative indices.

Ryans' observational report form (Ryans, 1960) is another illustration; this form has been factor analyzed, leading to the description of three major patterns of classroom behavior: understanding and friendly; responsible and businesslike; and stimulating and imaginative.

The third method, rating of teachers by students, has been less frequently employed, but useful instruments similar to those of Durflinger and Ryans are beginning to appear. Veldman and Peck (1963) have introduced a 38-item Pupil Observation Survey (POSR) which yields five factors: (*a*) friendly, cheerful; (*b*) knowledgeable, poised; (*c*) interesting, preferred; (*d*) strict control; and (*e*) democratic procedure. Factors I and IV seem to correlate highly with supervisors' ratings of the same teachers. Factor II has a moderate correlation, whereas Factors III and V bear little relationship.

The "criterion problem" is still far from solved, but as indicated above there do ap-

pear to be instruments and techniques available which can provide acceptable indices of performance.

If we turn from the criterion to consideration of the psychological tests which might be applied as predictors of teaching effectiveness, further problems are encountered. Much of the evidence is negative. For example, in the intellectual realm, Durflinger obtained correlations of −.08, −.13, and −.13 for the quantitative, linguistic, and total scores on the American Council on Education Psychological Examination (ACE) against supervisors' ratings of 150 student teachers on his 41-item scale. Stern (1963) concluded that intellectual measures have in general been unpredictive in this realm.

Several years ago, the Minnesota Teacher Attitude Inventory (Cook, Leeds, & Callis, 1951) seemed very promising because of the statistical and psychological sophistication invested in its development. Unfortunately, research evidence has not always sustained this promise. Sandgren and Schmidt (1956) classified 393 student teachers in high-middle-low categories on the MTAI, but these three categories bore no relationship to supervisors' evaluations. Durflinger, in the study already mentioned, found a correlation of −.12 between the MTAI and the criterion of teaching effectiveness, and Burkard (1965) did not find the MTAI to be valid.

The Minnesota Multiphasic Personality Inventory (Hathaway & McKinley, 1943) has been widely applied and because of its sensitivity to problems of maladjustment, neuroticism, and anxiety would seem to be relevant at least to *ineffectiveness* as a teacher. The research evidence, however, has not been encouraging. Tyler (1954) administered the MMPI to men enrolled in student teaching in secondary schools, and attempted to develop predictive indices on the basis of individual scales, combinations of scales, and item analysis of the inventory; these efforts were unsuccessful when checked in cross-validation.

Gough and Pemberton (1952) obtained moderately positive results in an attempt to forecast teaching effectiveness among female subjects (*S*s), using the "sign" approach on the MMPI, and Gowan and Gowan (1955) also obtained moderately favorable results in an item analysis of the inventory. However, Michaelis (1954), in a very searching study of the MMPI among 271 females in student teaching, obtained negative findings for both scale and item analyses, and Moore and Cole (1957), using the MMPI, were unable to distinguish among teachers rated as best, above average, average, below average, and poorest.

Michaelis also obtained relatively unpromising results with the Heston Personal Adjustment Inventory (Heston, 1949), the Minnesota Personality Scale (Darley & McNamara, 1941), and the Minnesota T-S-E Inventory (Evans & McConnell, 1942).

Perhaps the problem has been, as suggested by Peck (1960), that these inventories stress qualities which are not of great relevance to teaching, and hence are unable to achieve an adequate level of predictive accuracy. The California Psychological Inventory (Gough, 1957), developed to assess "folk dimensions" of interpersonal and interactional behavior, might be less subject to Peck's animadversion; that is, variables such as socialization, psychological-mindedness, and flexibility (all of which are scaled on the inventory) might be more relevant to teaching performance and therefore useful in its prediction.

Three studies employing the CPI to forecast teaching performance may be cited. The first, by Durflinger (1963), has already been mentioned. Against the criterion of ratings, the 18 scales gave individual correlations ranging from −.27 to +.12. In the second study (Hill, 1960), student teachers were dichotomously classified as better or poorer by faculty supervisors, and an effort was made to discriminate between the two groups by means of CPI scales. Some minor trends were observable (e.g., higher scores on the So scale for the higher-rated teachers), but the differentiations were far short of a level of practical utility.

Hill (1961) next conducted an item analysis of the inventory, pitting 50 higher-rated versus 50 lower-rated stu-

dents; 30 items significant at or beyond the .05 level of confidence were identified. These 30 items were scored on hold-out samples of 49 versus 52, giving rise to a point biserial correlation of +.19.

With such unencouraging results, a reader might well ask why one should continue to investigate the issue. The answer is that in spite of rather modest findings in these studies, taken separately, certain common trends were observable, leading to the notion that a pooling of data, and new analyses seeking patterns and combinations of variables, might yield more substantial returns. The remainder of this paper will report the outcome of an attempt to identify a predictive cluster of scores on the inventory.

FIRST ANALYSIS

Because the Durflinger criterion was in continuous form (supervisors' ratings on a 41-item schedule), it was decided to begin with *S*s from this study. The sample comprised 91 female students from the University of California, Santa Barbara, doing classroom teaching under supervision. Each *S* was rated by two different supervisors, one during the first semester of teaching and the other during a second semester; interjudge reliability of rating was found to be +.81. The combined score was used as a criterion.

Correlations of these ratings with the scales of the inventory were modest, ranging from −.18 for Fx (flexibility) and −.14 for Gi (good impression) up to +.13 for Sy (sociability) and So (socialization).

A search was therefore initiated for a combination of scales which might provide a more useful basis of prediction than any single scale taken alone. A stepwise multiple regression analysis of the 18 scales of the inventory against the teaching criterion was conducted, giving rise to a five-variable equation, including Sy, So, and Py (psychological-mindedness, with positive weights, and Gi and Fx with negative. Scores for the 91 *S*s, computed according to the equation, correlated +.36 with the criterion.[1]

The next step in the analysis was to cross-validate this preliminary equation on the sample of 202 *S*s drawn from Hill's investigations. These *S*s included 124 females and 78 males, tested and observed in the instructional program at Ball State University in Muncie, Indiana. The criterion of teaching effectiveness for these students was dichotomous, with 63 females and 35 males being classified by supervisors as superior ("A" group), and 61 females and 43 males as inferior ("B" group).

Table 1 presents means and standard deviations on the scales of the CPI for the higher-rated and lower-rated subsamples, and also the results of the t tests for significance. The higher-rated males scored significantly higher on the Py scale ($p < .01$), and significantly lower on Fx ($p < .05$). Higher-rated females exceeded lower-rated females on the Re (responsibility) and Ac (achievement via conformance) scales ($p < .01$ in both instances).

When the preliminary equation was used to compute scores for these 202 *S*s, higher-rated females had a mean of 50.37, *SD* 7.35, and lower-rated females a mean of 48.80, *SD* 7.95; the difference of 1.57 gave a t-ratio of 1.13 ($p = .26$) and a biserial correlation of +.13.

For males, the results were more encouraging, with a mean of 51.86, *SD* 8.31 for higher-rated *S*s versus 47.07, *SD* 6.34 for lower-rated; the difference of 4.79 gave a t ratio of 2.83 ($p < .01$), and a biserial correlation of +.37. These results on cross-validation of the preliminary equation, although modest, were considered sufficient to warrant further exploration.

Step 3 in the analysis was a sequential multiple regression analysis conducted over the entire sample of 293 *S*s. For this analysis, the 91 students from Durflinger's project were dichotomized into 46 higher-rated and 45 lower-rated, on the basis of higher and lower totals on the rating form. Correlations of each scale with the 2-versus-1 dichotomy were utilized in the analy-

[1] This analysis and the others reported in the paper were carried out at the University of California, Berkeley, Computer Center. The authors wish to thank the Center for granting computational time, and Quintin Welch and Susan Hopkin for conducting the analyses.

TABLE 1
COMPARISON OF HIGHER-RATED AND LOWER-RATED STUDENT TEACHERS IN THE BALL STATE UNIVERSITY SAMPLES

Scale	Males					Females				
	Higher		Lower			Higher		Lower		
	M	*SD*	*M*	*SD*	diff	*M*	*SD*	*M*	*SD*	diff
Do	29.80	4.80	30.42	5.78	−0.62	28.49	5.96	26.92	5.31	1.57
Cs	19.86	2.98	19.93	3.35	−0.07	21.48	3.17	20.95	3.73	0.53
Sy	25.83	4.87	25.56	4.83	0.27	26.38	4.98	25.48	4.49	0.90
Sp	37.71	5.47	37.05	5.58	0.66	35.35	5.44	35.05	5.20	0.30
Sa	22.57	3.48	23.33	3.17	−0.76	22.68	3.61	22.90	2.95	−0.22
Wb	39.29	2.74	38.42	4.12	0.87	38.44	4.15	37.03	5.06	1.41
Re	31.97	4.06	32.42	3.61	−0.45	34.79	2.44	32.93	3.97	1.86**
So	39.80	5.52	37.53	4.90	2.27	41.46	4.67	40.23	5.51	1.23
Sc	30.49	7.10	29.91	6.49	0.58	33.48	6.28	30.92	8.29	2.56
To	23.74	4.12	22.95	3.62	0.79	25.37	3.81	23.98	4.87	1.39
Gi	18.71	6.18	18.98	5.84	−0.27	19.33	5.41	18.13	7.29	1.20
Cm	25.89	1.68	25.56	2.75	0.33	25.84	2.87	25.33	2.71	0.51
Ac	29.97	4.14	29.00	4.08	0.97	30.49	3.15	28.46	4.00	2.03**
Ai	20.14	3.90	19.67	2.97	0.47	21.75	3.07	21.02	3.96	0.73
Ie	40.17	4.11	39.72	4.53	0.45	41.24	4.43	39.21	5.96	2.03*
Py	12.51	2.77	10.79	2.85	1.72**	11.38	2.71	11.10	3.02	0.28
Fx	9.26	3.44	9.74	3.92	−0.48*	10.08	2.82	10.51	3.47	−0.43
Fe	15.40	3.56	17.51	3.79	−2.11	23.51	3.38	24.07	3.53	−0.56

Note. For males, $N = 35$ for higher, $N = 43$ for lower; for females, $N = 63$ for higher, $N = 61$ for lower.
* $p < .05$.
** $p < .01$.

sis. The equation derived from the regression analysis is offered below:

$$\text{Teaching effectiveness} = 14.743 + .334\text{So} - .670\text{Gi} + .997\text{Ac} + .909\text{Py} - .446\text{Fx}.$$

The beta weights are for use with raw scores on the five CPI scales, and the constant of 14.743 has been set so that the mean of an array of computed scores will converge on 50.00.

This equation was then used to compute scores for each of the 293 students from the two programs. Table 2 presents a summary of the results obtained from an analysis of these scores. All three differentiations are significant, as of course would be expected for application of an equation to the cases on which it was developed.

TABLE 2
RELATIONSHIP BETWEEN SCORES ON THE EQUATION FOR TEACHING EFFECTIVENESS AND CRITERION RATINGS IN THE SAMPLES USED TO DERIVE THE EQUATION

Samples	*N*	*M*	*SD*	diff	*t*	r_{bis}
University of California						
High-rated females	46	53.11	3.20	2.00	2.93*	.37
Low-rated females	45	51.11	3.23			
Ball State University						
High-rated females	63	51.95	4.36	2.16	2.50*	.28
Low-rated females	61	49.79	5.15			
High-rated males	35	52.60	5.68	3.60	2.92*	.40
Low-rated males	43	49.00	5.05			

* $p < .01$.

CROSS-VALIDATION

The essential step for any equation such as this is cross-validation. About the time we were completing our analyses to this point, a paper by Veldman and Kelly (1965) was published in which CPI data were presented; these two authors have been kind enough to permit cross-validation of the equation on their sample.

The *Ss* studied by Veldman and Kelly were 34 University of Texas women. The sample was dichotomized by supervisors into two categories, more effective versus

less effective, and these classifications were confirmed by interpretation of scores on the Peck Incomplete Sentences Test for teaching potential (Peck, 1960).

CPI scores were computed for each girl, using the formula given above. For the 17 higher-rated student teachers, a mean of 52.41, *SD* 5.17 was obtained, and for the 17 lower-rated the mean was 47.94, *SD* 6.59. The difference of 4.47 produced a t ratio of 4.27 ($p < .001$), and a biserial correlation of +.44. Although the sample of 34 is too small to permit broad generalization, the results tend clearly to confirm the validity of the equation.

The analyses reported to this juncture all deal with magnitudes and differences. An equally important question concerns the percentage of error which would occur if scores on the equation were used to classify individuals as higher- or lower-rated.

The proper cutting point to use in such classification is that which most closely approximates the split given by the criterion dichotomies. The score of 52 best met this requirement. Therefore, students with scores of 52 or above were classified as "highs" and those with scores of 51 or below as "lows." If a student is called high by the test and is also high on the criterion dichotomy, then he may be termed a "hit." Likewise, if a student is called low by the test, and is also low on the criterion, his classification is correct. An error in either direction may be called a "miss."

The question now becomes, what is the percentage of "hits," or the batting average, if we use the equation to forecast high versus low on the criterion? For the females from Santa Barbara, 60 (65.9%) of the 91 *S*s are correctly classified. For the 124 girls from Muncie, 70 (56.5%) were correctly designated. For Hill's 78 males, 52 (66.7%) were properly identified. And for the Veldman-Kelly *S*s, the only sample which is fully independent for purposes of cross-validation, 23 students (64.7%) were correctly categorized.

We must also compute the chance level for such classification, for unless the criterion split is precisely 50–50 the chance level of accuracy will be greater than 50%; this occurs because the "best guess," when the frequencies depart from a 50–50 basis, is that any individual will belong to the larger category. The chance base lines for the four samples in the preceding paragraph are as follows: Durflinger females, 50.6%; Hill females, 50.8%; Hill males, 55.1%; and Veldman-Kelly females, 50.0%.

Conceptual Analysis

Having demonstrated that the CPI equation for student teaching effectiveness has at least moderate validity, the next question to raise is this: "What kind of an individual is it, in everyday language and description, who is identified as a good prospect by this equation?" We wish to turn, in other words, from a consideration of the *predictive validity* of the equation to a study of its *diagnostic implications.*

The methodology to be used in answering this question is one which has been called "conceptual analysis [Gough, 1965]." Its essential feature involves study of individuals rated higher or lower by the equation, and from identifying their prominent characteristics to infer the underlying psychological dimensionality of the measure.

Two research samples were available for carrying out this conceptual analysis. The first was composed of 101 college males, members of three different fraternities, at the University of California, Berkeley. Each boy had taken the CPI and each had also been described, by five of his fellow-members, on the Adjective Check List (ACL; Gough & Heilbrun, 1965). By summing the number of times a word was checked about a boy by these five peers, a descriptive total was obtained for each of the 300 words in the instrument. Within each of the three subsamples, these 300 totals were converted to standard scores, so that the three subsamples could be combined into the one large sample of 101 *S*s.

From the CPI protocols, a student teaching effectiveness score was also computed for each boy, using the equation cited above. Then, this CPI score was correlated with the 300 peer descriptions gathered by

means of the ACL. A significant positive correlation between the CPI score and a word on the ACL would mean that this word was used in a differential way to describe *S*s scoring high on the equation; a significant negative correlation would mean that the word was used in a differential way to describe boys with low scores on the question. By accumulating the words with highest positive and largest negative correlations, a verbal portrait of the high-scoring and low-scoring boy can tentatively be drawn.

For girls, a similar sample of 92 *S*s from two sororities at the University of California, Berkeley, was available.[2] Here too, each girl had taken the CPI, and had been described on the ACL by five of her peers. Within each subsample, the 300 ACL totals were converted to standard scores so that the correlations between the equation and the 300 words could be computed on the full sample of 92 *S*s.

For the males, more than 35 correlations were significant at the .01 level of confidence (3 would be expected by chance), so the pattern of relationships between peer descriptions and the CPI index appears to be reliable. To clarify the findings and render the patterns more easily discernible, the 12 words with highest positive correlations will be listed first, and then the 12 words with largest negative correlations.

The 12 words used most typically to describe college males scoring high on the CPI equation were these (the words are listed in order of magnitude of correlation):

conscientious
practical
rational
moderate
methodical
planful
responsible
logical
reasonable
capable
thorough
reserved

The correlations range from a low of +.29 (for "reserved") to a high of +.37 (for "conscientious").

What does this set of 12 descriptions suggest about the high-scorer? He seems to be a diligent, effective individual, well-organized, attentive to the practical demands of his work, and thorough and conscientious in carrying out his duties. In temperament he is self-disciplined and reserved, not at all flamboyant or unconventional. He is the kind of person who can be counted on to display discretion and good judgment in any situation.

What about the low-scoring male on the equation? The 12 words used most differentially to describe him were these:

reckless
daring
pleasure-seeking
spendthrift
irresponsible
flirtatious
show-off
spontaneous
adventurous
michievous
quick
careless

A "syndrome" of temperament and behavior seems clearly evident in this cluster of descriptions. The low man on the CPI equation for forecasting teaching effectiveness appears to be undercontrolled, unbridled, too much dominated by his own impulses. Although in many ways an attractive personality (spontaneous, adventurous, quick), and probably original in his perceptions and ideas, he is too irresponsible and too careless to perform effectively in a day-by-day classroom situation.

While pondering these characterological implications of the CPI equation, we should note some of the factors which are

[2] Use of samples of fraternity and sorority members should not bias the findings, as it has been shown in earlier work (Gough, 1965) that essentially equivalent diagnostic implications of test variables are obtained from samples differing significantly on age, educational level, and occupational status. Sex differences are important in conceptual analysis, and for this reason separate inquiry was undertaken for male and female *S*s.

not included: The equation does not appear in any way to rest on intellectual ability, charm, assertiveness, or other qualities which one might hypothesize as determiners of scores on the inventory. The two patterns (for high-scoring and low-scoring males) are not merely good and bad, but rather two patterns more or less compatible with the demands of the criterion.

Attention is now directed to the females. Independent analyses are required because of the common finding that the same variable will have different implications for males and females. We should not, therefore, expect to find that the CPI equation rests on the same psychodynamic basis for both sexes.

As with the males, more than 35 descriptions correlated at or beyond the .01 level of confidence, and to clarify the relationships only the 12 words with highest positive and 12 with largest negative correlations will be listed. The 12 words used most differentially to describe high-scorers were these:

dominant
persevering
persistent
serious
opinionated
ambitious
demanding
logical
rigid
clear-thinking
determined
responsible

What was expected to happen did happen: The woman identified by the equation as a potentially effective student teacher is rather different from the man so identified. The high-scoring young lady is a strong and resourceful individual, clear and explicit about her goals, and resolute in pursuing them. In fact, her seriousness of purpose and determination are such that those who know her well find her somewhat rigid and opinionated, however worthy her ambitions and steadfastness.

The pattern has a touch of the negative in it, but one cannot conclude that *S* would be domineering or authoritarian in class, nor can one assume that pupils would find her objectionable. Retracing the chain of relationships which led to the development of the equation, we must recall that the original criterion consisted of ratings of effectiveness by supervisors thoroughly grounded in education philosophy and sensitive to any hint of autocratic or manipulative behavior.

The effective woman teacher, the adjectival analyses therefore suggest, may be one who, although single-minded and inexorable in her resolve, can nonetheless deal with her students in an insightful and responsible manner. Her personal friends may see a touch of inflexibility and dogmatism in her beliefs, but her students may experience these qualities as decisiveness and clarity.

Finally, what of the low-scoring college woman, how is she described by her friends and peers? The 12 adjectives most differentially applied were these:

curious
affectionate
careless
easy going
unconventional
dreamy
understanding
irresponsible
cheerful
natural
individualistic
thoughtful

Some of these words duplicate descriptions found for the low-scoring males (e.g., careless and irresponsible), but the flavor of the cluster is different. The low-scoring female is somewhat undercontrolled, to be sure, but she is nonetheless affectionate, thoughtful, and of an optimistic turn of mind. Hostility, aggression, rebelliousness,—all qualities which one might hypothesize as negatively related to teaching effectiveness—are alien to the pattern actually delineated. Our low-scoring *S* seems very likable, easy to get along with, a pleasant and undemanding individual.

But as a teacher she will not do; her lack of organization, overresponsiveness to distractions of the moment, and indifference to practical realities are drawbacks too great to be ignored.

The methodology of conceptual analysis, although simple in design and in application, is sufficiently contrary to "ordinary" procedure so that some readers may misunderstand what is being claimed for the interpretations sketched above. The four characterizations, it should be stressed, are not offered as descriptions of good and poor teachers, male and female. Rather, they seek to define four syndromes *which are diagnosed by the CPI equation*. If a male scores high on the equation, then it is likely that he will resemble the person described in the first portrait; and, we hasten to add, it is also likely that he will be an effective teacher and that supervisors will agree on his effectiveness.

If a female college student scores high on the CPI equation, she may be expected to be characterized by the constellation of traits and dispositions sketched in the third portrait. And, should she participate in student teaching, we can anticipate superior performance and high ratings by her supervisors.

The same admonitions hold for the portrayals of low-scorers on the equation: Low-scoring males will tend to be improvident, impulsive, and adventurous individuals—and poor candidates for teaching training; low-scoring females will tend to be impractical and undependable, albeit charming, and also poor risks for such training. There are many, perhaps innumerable, ways of being a poor teacher; the function of the CPI equation is to identify two important routes (one for each sex), and the purpose of conceptual analysis is to specify their personological parameters.

REFERENCES

BARR, A. S. Problems associated with the measurement and prediction of teacher success. *Journal of Educational Research*, 1958, **51**, 695–699.

BIDDLE, B. J., & ELLENA, W. J. *Contemporary research on teacher effectiveness*. New York: Holt, Rinehart & Winston, 1964.

BURKARD, M. I. Effectiveness of the MTAI in a parochial school setting. *Journal of Experimental Education*, 1965, **33**, 225–229.

CHAUNCEY, H., & DOBBIN, J. E. *Testing: Its place in education today*. New York: Harper & Row, 1964.

COOK, W. W., LEEDS, C. H., & CALLIS, R. *The Minnesota Teacher Attitude Inventory*. New York: Psychological Corporation, 1951.

DARLEY, J. G., & MCNAMARA, W. J. *Minnesota Personality Scale*. New York: Psychological Corporation, 1941.

DURFLINGER, G. W. Personality correlates of success in student teaching. *Educational and Psychological Measurement*, 1963, **23**, 383–390.

ELLENA, W. J., STEVENSON, M., & WEBB, H. V. (Eds.), *Who's a good teacher?* Washington, D. C.: American Association of School Administrators, 1961.

EVANS, C., & MCCONNELL, T. R. *The Minnesota T-S-E Inventory*. Chicago: Science Research Associates, 1942.

GAGE, N. L. (Ed.) *Handbook of research on teaching*. Chicago: Rand McNally, 1963.

GETZELS, J. W., & JACKSON, P. W. The teacher's personality and characteristics. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally, 1963. Pp. 506–582.

GOUGH, H. G. *Manual for the California Psychological Inventory* (Revised, 1964). Palo Alto, California: Consulting Psychologists Press, 1957.

GOUGH, H. G. Conceptual analysis of psychological test scores and other diagnostic variables. *Journal of Abnormal Psychology*, 1965, **70**, 294–302.

GOUGH, H. G., & HEILBRUN, A. B., JR. *The Adjective Check List manual*. Palo Alto, California: Consulting Psychologists Press, 1965.

GOUGH, H. G., & PEMBERTON, W. H. Personality characteristics related to success in practice teaching. *Journal of Applied Psychology*, 1952, **36**, 307–309.

GOWAN, J. C., & GOWAN, M. S. A teacher prognosis scale for the MMPI. *Journal of Educational Research*, 1955, **49**, 1–12.

HATHAWAY, S. R., & MCKINLEY, J. C. *Minnesota Multiphasic Personality Inventory manual*. (Revised 1951). New York: Psychological Corporation, 1943.

HESTON, J. C. *Heston Personal Adjustment Inventory—manual*. New York: World Book, 1949.

HILL, R. E., JR. Dichotomous prediction of student teaching excellence employing selected CPI scales. *Journal of Educational Research*, 1960, **53**, 349–351.

HILL, R. E., JR. An investigation of the California Psychological Inventory empirically keyed for dichotomous prediction of student teacher excellence. In E. M. Huddleston (Ed.), *The eighteenth yearbook of the National Council on Measurements Used in Education*. Ames, Iowa: N.C.M.U.E., 1961.

MICHAELIS, J. U. The prediction of student teaching from personality and attitude inventories.

University of California Publications in Education, 1954, **11,** 415–484.

MOORE, C. H., & COLE, D. The relationship of MMPI scores to practice teaching ratings. *Journal of Educational Research,* 1957, **50,** 711–716.

PECK, R. F. Personality patterns of prospective teachers. *Journal of Experimental Education,* 1960, **29,** 169–175.

RYANS, D. G. *Characteristics of teachers.* Washington, D. C.: American Council on Education, 1960.

SANDGREN, D. L., & SCHMIDT, L. G. Does practice teaching change attitudes toward teaching? *Journal of Educational Research,* 1956, **49,** 673–680.

STERN, G. G. Measuring noncognitive variables in research in teaching. In N. L. Gage (Ed.), *Handbook of research on teaching.* Chicago: Rand McNally, 1963. Pp. 398–447.

TYLER, F. T. The prediction of student-teaching success from personality inventories. *University of California Publications in Education,* 1954, **11,** 233–313.

VELDMAN, D. J., & KELLY, F. J. Personality correlates of a composite criterion of teaching effectiveness. *Alberta Journal of Educational Research,* 1965, **11,** 102–107.

VELDMAN, D. J., & PECK, R. F. Student-teacher characteristics from the pupil's viewpoint. *Journal of Educational Psychology,* 1963, **54,** 346–355.

(Received April 3, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 2, 128-132

TEACHERS' RATINGS OF STUDENT PERSONALITY TRAITS AS THEY RELATE TO IQ AND SOCIAL DESIRABILITY[1]

JAMES W. BARNARD
George Peabody College

PHILIP G. ZIMBARDO
New York University

AND SEYMOUR B. SARASON
Yale University

4 groups of 2nd- and 3rd-grade school children varying in their IQ status and their test anxiety status were rated by their teachers on a series of 24 personality and school performance characteristics. No relationships were found between the teachers' ratings of their children on these characteristics and the childrens' test anxiety status. Significant relationships were found between the teachers' ratings on 14 of the 24 characteristics and the childrens' IQ status. It was further found that there existed a relationship between the extent to which a characteristic was judged to be desirable and the extent to which the item differentiated between high- and low-IQ children, in the direction of the desirable characteristics being more frequently attributed to the high-IQ children. Some evidence was presented that supported the position that these evaluations reflected, at least in part, the biases of the teacher raters rather than simply actual behavioral differences of the children.

In most educational systems it is expected that a teacher be aware of psychological differences that exist among his students. In some instances the process of evaluation of nonacademic behavioral traits has grown to the point where the modern elementary school teacher must be as attentive to a student's entire personality development as he is to his intellectual growth. Thus, a student's permanent school record now often contains teachers' assessments which, when dressed in the psychologist's jargon, appear under such headings as "extent of peer group dependence," "control of aggression," "introversion," "anxiety," and so forth. These assessments are used not only in predicting a student's performance within the elementary school setting, but as part of his permanent record play an important role even in the screening of applicants for high school and college awards and fellowships, as well as for admission to graduate and professional schools.

Moreover, it should be recognized that the formal categorization of student personality traits by teachers can exert not only a controlling influence on how that particular teacher then perceives, organizes, and interprets later behavior, but as part of the student's record, it may also establish the frames of reference through which subsequent teachers view the student. The question of how well teachers can perform such an evaluative function is obviously of first-rate importance.

The present study represents an attempt to investigate the ability of elementary school teachers to discriminate among their students on relevant and important personality variables. Further, an attempt was made to investigate the relationship between teachers' ratings of their students on these personality variables and the students' IQ test scores.

METHOD

Subject Selection

All of the second- and third-grade school children of Hamden, Connecticut, were administered the Test Anxiety Scale for Children (TASC) and the Lorge-Thorndike intelligence test as a part of a longitudinal study being carried out at Yale University (Sarason, Hill, & Zimbardo, 1965). From this population, those children scoring in the upper and lower fifteenth percentile of the anxiety distribution ($N = 320$) were selected to form the high-anxiety (HA) and low-anxiety (LA) groups, respectively. Within each of these groups subjects

[1] This study was financed by a grant to Yale University (Seymour B. Sarason, principal investigator) from the National Institute of Mental Health.

(*Ss*) were matched on IQ score as closely as possible in order to obtain subgroups of relatively high IQ (HIQ) and low IQ (LIQ). The effectiveness of this matching is evident from the group mean IQ of 115 for the HA-HIQ group and 116 for the LA-HIQ, as well as from the means of 99 for the HA-LIQ and 97 for the LA-LIQ groups. A total of 96 *Ss* were thus finally chosen, 24 in each of these four experimental groups.

Teachers' Ratings

These 96 students were rated by their classroom teachers ($N = 54$) on 24 personality and school performance characteristics. These characteristics were presented in pairs of contrasting trait names, along with a working definition of each of the terms in the pair. Each teacher had to decide first which of the two terms most accurately described the child, then she had to determine by use of a 5-point scale the degree to which the child approached the extreme description given for that trait. Each teacher was given a chance to discuss this rating task with an experimenter to insure that she understood what was being asked of her. After this discussion, each teacher worked on the task privately on her own time. The complete list of traits is given in Table 1.

The obtained ratings were then summed across *Ss* within each group for each of the 24 traits separately, and then subjected to simple between-*Ss* analyses of variance with test anxiety level and IQ group as main effects.

TABLE 1

TEACHERS' RATING SCALE OF STUDENT CHARACTERISTICS

Trait No.
1. Anxious: Unanxious
2. (Dependent: Independent)
3. Shows or expresses emotions: Hides or suppresses emotions
4. (Communicates easily: Difficulty communicating)
5. Aggressive: Submissive
6. Impulsive: Cautious
7. Sensitive: Not sensitive
8. Tense: Relaxed
9. (Ambitious: Unambitious)
10. (Adapts to changes: Set in ways)
11. (Well-liked: not well-liked)
12. (Mature psychologically or emotionally: Immature psychologically or emotionally)
13. Withdraws: Sociable
14. Daydreams: Does not daydream
15. Active: Inactive
16. Overachievers: Underachievers
17. (Learns slowly (new material): Learns quickly (new material))
18. (Retains material: Forgets material)
19. Fears failure: Does not fear failure
20. (Pays attention: Does not pay attention)
21. (Strong conscience: Weak conscience)
22. (Feminine: Masculine)
23. (Pessimistic: Optimistic)
24. (Responsible: Not responsible)

Note.—Trait numbers in parentheses have highest agreement as to the desirability of that trait, while those not in parentheses have least agreement, and thus are least clearly positive or negative.

Desirability of Trait Scoring

A second group of judges was used to provide additional information about the social desirability of each of the traits used in the teacher rating schedule. These data were necessary to test an hypothesis which emerged after preliminary analysis of the ratings. Fifteen teachers in the Yale Master of Arts in Teaching program (M.A.T.), all of whom had some teaching experience, were asked to judge how desirable it was for an elementary school child to exhibit each one of the bipolar traits which defined the 24 trait dimensions.

These judgments were made on 10-point scales and were computed in such a way that high scores for a given item would indicate agreement among the judges as to its desirability, while low scores would indicate disagreement or lack of clarity about the desirableness of that item. The items were presented to the judges using the same format as was used in the presentation to the teachers when they rated their own students.

The relationship between how conceptually clear the desirability of a trait was and how well it discriminated on the teacher ratings between IQ groups was established by means of correlation. The correlation coefficient obtained was between the mean desirability score of a trait as determined by the M.A.T. judges, and the difference in the total teacher rating score between high- and low-IQ groups on that trait. Thus, a high positive correlation would indicate that the better a trait differentiated between high- and low-IQ *Ss*, the greater the agreement among judges as to its desirability.

RESULTS

The teachers did not differentiate in their ratings between LA and HA children on any of the 24 traits. The nonsignificant differences between anxiety groups can be seen from the F values presented in Table 2.

On the other hand, it is obvious from the rest of the evidence presented in this table that the teachers did discriminate on the basis of IQ. On 14 of 24 traits, students with high IQ were characterized differently from those with low IQ at beyond the .05 level of confidence. Over the com-

TABLE 2

F-Values of Anxiety and IQ Sources of Variance for Each Trait

Trait	Anxiety	IQ
1	.84	1.30
2	.07	16.39**
3	.00	1.98
4	.43	.77
5	1.00	5.70*
6	3.25	2.97
7	1.29	6.34*
8	.21	.75
9	.25	10.97**
10	.98	2.43
11	.23	4.87*
12	.49	7.86**
13	1.83	12.28**
14	.03	4.74*
15	.73	8.20**
16	2.12	5.43*
17	.07	28.38**
18	.03	13.84**
19	.01	1.73
20	.67	7.20**
21	2.27	2.79
22	.08	.45
23	.37	3.36
24	.01	7.34**

Note.—Trait numbers correspond to traits listed in Table 1.
* $p < .05$.
** $p < .01$.

bined teacher ratings the difference between HIQ and LIQ *S*s was highly significant ($F = 20.03, p < .001$).

Although none of the second-order effects was significant, nevertheless, in most instances the differences in teacher ratings between *S*s of the two IQ groups were greater among HA *S*s than among LA ones. Thus, LA children tended to be rated more similarly, whether their IQ was high or low, than were HA children. To substantiate this observation, a difference score for each trait was generated by subtracting the scores of the high-IQ *S*s from the low-IQ *S*s within each of the anxiety groups. The analysis of these difference scores, using a *t* test with nonindependent observations, demonstrated that difference scores on these 24 traits between IQ subgroups within the HA group were significantly greater than those difference scores of the IQ subgroups within the LA group ($t = 2.45, p < .05$).

A further examination of the data revealed that there were marked differences in how well various traits differentiated high- from low-IQ *S*s. It appeared that those items that showed the largest differences in scores between the two IQ groups were the items that were most clearly desirable traits for an elementary school child to have. Apparently, then, this inconsistency in the ability of items to differentiate between high and low *S*s was related to a tendency on the part of the teachers to assign more favorable ratings to high-IQ *S*s to the extent that it was clear to them what a favorable rating would be. To investigate this notion a correlation was computed between IQ group differences on each trait and the desirability score of that trait as judged by the M.A.T. teachers.

The traits in parentheses in Table 1 are those that demonstrated the highest clarity of desirability (i.e., above the median), while for the others there was a lack of agreement among judges as to whether one of the bipolar traits was desirable and the other one undesirable for a child to exhibit in school. The product-moment correlation between item discriminability and desirability was .60 ($p < .01$).

Discussion

The results of this study can be interpreted in such a way as to cast some doubt on the validity of ratings by elementary school teachers of student personality traits. The present teachers were unable to distinguish reliably between students who were extremely different, by their own self report, in test anxiety, on traits that have been found in past research to be related to the variable of test anxiety. Although other interpretations are possible, these data are consistent with the earlier conclusion drawn from the research of Sarason.

In none of our studies using teachers' ratings of

anxiety in relation to a test performance or a child's self-report criterion is there evidence that teachers can recognize the anxious child to a degree which would be of practical significance [Sarason, Davidson, Lighthall, Waite, & Ruebush, 1960, p. 265].

It does appear from our data, however, that teachers are sensitive to differences in IQ level. This sensitivity is reflected in their differential rating of children with high- and low-IQ scores on a wide variety of academic and personality traits. The child with a high IQ tends to be perceived, relative to a child with a low IQ, as one who learns quickly, pays attention, retains material, overachieves, and is ambitious. Such traits are the obvious correlates of high IQ for the adequately motivated student. However, even on traits which bear little correspondence to academic performance and intellectual functioning, teachers discriminate between children of different intelligence levels. The bright child tends to be seen by the teachers as being less dependent and daydreamy and more aggressive, while at the same time being more sensitive, mature, sociable, popular, and active.

The issue which becomes immediately apparent is whether these evaluations are reflecting actually occurring behavioral differences, or whether they are distortions of social reality. Without an independent criterion analysis of each of these traits, there cannot be an unequivocal answer to such a problem. However, several lines of converging evidence lead us to believe that a major source of variance in these ratings is accounted for by teacher bias in perception as a consequence of overly positive evaluations of bright children. First, it was demonstrated that these teachers distinguished most clearly between IQ levels on those traits which could be most easily categorized as desirable or undesirable for a child to possess or exhibit in school. Second, it was learned (after completion of our data collection) that all teachers in the sample had access to, and were familiar with, information pertaining to the results of the IQ and achievement testing which were routinely conducted in this school system. Finally, a biasing in terms of a halo effect appears to be a tenable explanation for this data, since teachers evaluated anxious children who were bright differently from anxious children who were not, and did so on traits shown by previous research *not* to be characteristic of the bright, anxious child. Thus, for example, while these teachers characterized the bright, anxious child with the desirable traits of "independence" and "adaptability," it has been demonstrated that the bright but anxious child is extremely dependent upon task and instruction factors as well as upon the approval of authority figures, all of which inhibit spontaneity, independence, personal expression, and flexibility in school settings (Sarason et al., 1960).

It is also interesting to note that teachers do not consistently differentiate in their ratings of the bright and the nonbright children when these students have low levels of test anxiety. In some way then, the bright, anxious child is perceived as special and possessing traits of which teachers approve. In short, teachers are most positive about such children.

It is likely, however, that this favorable attitude is engendered in large part by the child's dependent need for approval by the teacher and by his attempts to secure it. By not recognizing that such a child is frequently experiencing anxiety in relation to school and the resultant evaluation of his abilities (the definition implicit in the test anxiety construct), teachers are unable to provide the help necessary to improve the child's selfconception, and in fact may even reinforce these test anxiety attitudes. In turn, these attitudes may generalize to influence many kinds of behavior (e.g., speech, as shown by Zimbardo, Barnard, & Berkowitz, 1963) and become the core of an enduring personality syndrome. Such a pattern of attitudes may cause many of the bright but anxious students who do get into college to lower their levels of aspiration, and become satisfied with minimal standards of learn-

ing which do not require utilization of their full intellectual and creative capacity (Mandler & Sarason, 1952).

The results of this study are consistent with a recent plea for teacher training in the knowledge and use of psychological variables, as well as in detection of subtle stimulus cues necessary for the understanding, modification, and control of behavior (Sarason, Davidson, & Blatt, 1962). The questions raised for future research are the extent to which knowledge of a child's IQ or anxiety level influences the overt classroom behavior of teachers in their handling of their students, and to what extent is this recognition perceived and reacted to by the students.

REFERENCES

MANDLER, G., & SARASON, S. B. A study of anxiety and learning. *Journal of Abnormal and Social Psychology*, 1952, **47**, 166–173.

SARASON, S. B., DAVIDSON, K. S., LIGHTHALL, F. F., WAITE, R. R., & RUEBUSH, B. K. *Anxiety in elementary school children.* New York: Wiley, 1960.

SARASON, S. B., DAVIDSON, K. S., & BLATT, B. *The preparation of teachers.* New York: Wiley, 1962.

SARASON, S. B., HILL, K., & ZIMBARDO, P. G. A longitudinal study of the relation of test anxiety to performance on intelligence and achievement tests. *Monographs of the Society for Research in Child Development*, 1965, **29**(7, Whole No. 98).

ZIMBARDO, P. G., BARNARD, J. W., & BERKOWITZ, L. The role of anxiety and defensiveness in children's verbal behavior. *Journal of Personality*, 1963, **31**, 79–96.

(Received April 10, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 2, 133-138

COGNITIVE TRANSFER IN VERBAL LEARNING: II. TRANSFER EFFECTS AFTER PREFAMILIARIZATION WITH INTEGRATED VERSUS PARTIALLY INTEGRATED VERBAL-PERCEPTUAL STRUCTURES

JAMES H. REYNOLDS[1]
Colgate University

A previous experiment (Reynolds, 1966) showed that prefamiliarization of verbal items embedded in an integrated and meaningful map structure produced positive transfer to the learning of sentences related to the map configuration. The present study extended the initial findings by showing (a) that prefamiliarization with a single, integrated map structure produced greater transfer to sentence learning than did prefamiliarization with the same map which had been fragmented into separate and discrete pictures; (b) that the positive transfer effect can be obtained after a 10-min. rest interval between the familiarization and the learning tasks; and (c) that the effect can be obtained with differing materials, testing methods, and age groups. The results were interpreted as evidence that the transfer observed was due to the wholeness or completeness of the map structure, and that both the first and the present experiment demonstrate a stable cognitive mechanism which facilitates rote learning of sentences.

Previous research (Reynolds, 1966) has shown that prefamiliarization of verbal stimuli embedded in an integrated and meaningful pictorial map produced positive transfer to the learning of simple sentences which contained factual material related to the previously-studied map. The transfer obtained was significantly greater than that obtained for any of five control conditions which received prefamiliarization with the verbal stimuli, the pictorial map, or nonintegrated combinations of each. It was concluded that the integration of the verbal and perceptual stimuli into a single meaningful structure, rather than familiarization with these components separately, was responsible for the positive transfer observed. The theoretical explanation of this result was that the integration of verbal and pictorial stimuli into a single meaningful whole permitted the formation of an assumed mental organization akin to Tolman's (1948) concept of a "cognitive map," which persisted in memory and aided later rote learning of the related sentences.

The present experiment attempted to explore further the effects of a verbal-perceptual structure upon transfer to a rote verbal task. The main problem investigated was whether the positive transfer obtained in the previous experiment was due to the *wholeness* or *completeness* of the verbal-perceptual stimulus configuration, or to the pairing of two types of stimuli—letters and pictures. A cognitive interpretation would assume that the wholeness of the perceptual configuration (i.e., the integrated map) provided a structure in which each part was meaningfully related, and thus could be recalled and utilized at the time of the learning task. Alternatively, it is possible that the pictorial characteristics of the map simply provided discrete perceptual stimuli which, added to the discrete verbal stimuli, provided stronger but still not necessarily integrated stimulus learning at the time of prefamiliarization. The latter alternative permits an S-R interpretation of the previous results by stating that the individual verbal stimuli were learned better when associated with discrete picture components during prefamiliarization than when presented alone, and thus the presence of discrete pictorial stimuli—and not the wholeness of the total map structure—was responsible for the transfer observed.

To evaluate these alternatives, the

[1] This research was supported by a research grant from the United States Office of Education, Contract No. OE-6-10-43.

present experiment compared groups which received prefamiliarization with a single, integrated map structure with other groups receiving the same verbal-pictorial stimulus combinations but fragmented so that the subject (S) was exposed to a series of separate word-picture stimuli rather than an integrated and meaningful whole. Controls received only verbal stimuli during prefamiliarization. The main hypothesis, in accordance with a cognitive theory, was that the integrated-map groups would demonstrate greater transfer to a related sentence-learning task than would the groups familiarized with the discrete word-picture materials.

A second problem investigated dealt with retention of prefamiliarized cognitive material. In the initial experiment, the transfer learning task was presented immediately following the prefamiliarization period. It is possible that, although the cognitive-map treatment yielded superior transfer in this immediate-memory situation, its positive effect may have been due to short-term memory of the preceding stimulus configuration rather than to the presence of a stable cognitive structure built up during the prefamiliarization period. Were this the case, the effect might be expected to dissipate over a short rest interval, in a manner similar to that demonstrated by Peterson and Peterson (1959) and others using different types of stimulus material. To test this possibility, in the present study a 10-minute rest interval was inserted between the end of the prefamiliarization period and the beginning of the transfer learning task.

Finally, an attempt was made to test the generality of the previous findings by using *Ss* from a different age group, employing two differing pictorial-map configurations, and administering the learning tasks at two different presentation rates by a study-test, rather than a free-recall, method.

METHOD

Subjects

The *Ss* were 36 boys and 36 girls between the ages of 15 and 18, all enrolled in a summer special-study program conducted by Colgate University for high school students of above-average ability. All *Ss* volunteered, and were paid a nominal sum for participation. For each sex, *Ss* were assigned to one of the six experimental conditions according to a predetermined order as they appeared individually for the experiment. The assignment method provided equal numbers of boys and girls in each condition.

Design

The general design was similar to that described in detail in Reynolds (1966). In the first stage of the experiment, map sketches of varying degrees of meaningful structure were presented for learning to *Ss* in the different groups. Following this first, or prefamiliarization, stage all *Ss* received a 10-minute rest interval during which they worked on a puzzle which was unrelated to the experimental tasks. In Stage 2, all *Ss* received a rote sentence-learning task in which the sentences to be learned were related to the structure materials used in Stage 1. The main hypothesis was that a meaningful and integrated map structure presented in Stage 1 would provide S with a cognitive structure which would be retained over the 10-minute rest interval and would transfer positively to the Stage 2 learning task, whereas *Ss* given Stage 1 maps which were not integrated into a single meaningful context would fail to form and maintain a cognitive structure and thus would demonstrate less positive transfer to the Stage 2 task.

Materials and Procedure

Stage 1. Three types of materials, designating three levels of cognitive structure, were used in Stage 1. The Cognitive Group received for study an 8½ × 11-inch map sketch depicting a common scene which contained eight parts. Each of the eight parts was labeled with a consonant-vowel-consonant (CVC) of 85–100% association value (Glaze, 1928). At a second level, the Picture Group received for study an 8½ × 11-inch sheet on which were depicted the same eight labeled parts making up the map presented to the Cognitive Group; but these parts were separated from each other by borders, so that the sheet contained eight separate pictures with CVC labels rather than a single integrated scene containing eight meaningfully related parts. At a third level, the Label Group received for study in Stage 1 an 8½ × 11-inch sheet on which were printed the eight CVC labels but which contained no pictures.

At each of these three levels of structure, two types of task stimuli were employed. Half of the *Ss* in the Cognitive Group were presented with the same map used in the previous experiment, and shown in Figure 1a. This map, designated the crossroads (CR) map, showed a modern highway intersection and its surroundings, including eight labeled parts: an airstrip, gas station, shopping center, diner, farm, train, trailer-truck, and patrol car. The other half of the *Ss* in the Cognitive

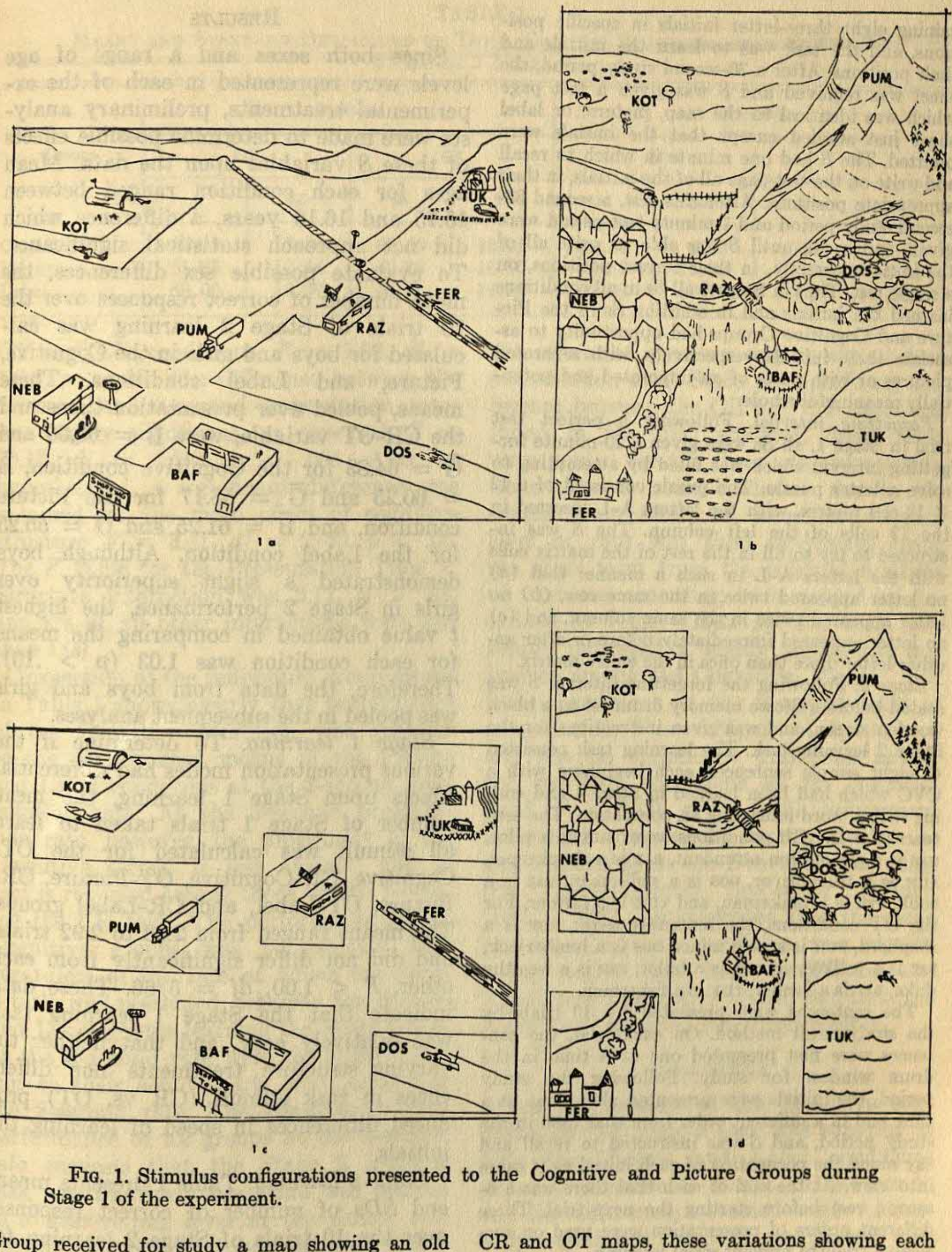


FIG. 1. Stimulus configurations presented to the Cognitive and Picture Groups during Stage 1 of the experiment.

Group received for study a map showing an old town (OT) and its surroundings. This map, shown in Figure 1b, contained CVC labels embedded at eight points in the total scene: in the town, on a flat plain above town, in the mountains, in the forest, on a lake, by a farm, by a bridge, and by a large castle-like building. As Figures 1c and 1d illustrate, the Picture Groups for the CR and OT conditions received fragmented variations of the CR and OT maps, these variations showing each labeled part in the same relative position as it appears on the integrated map, but separated from other parts by a border and blank space. The CR and OT Label Groups received white sheets of paper on which were printed CVCs in the same relative positions as shown in Figures 1a and 1b, respectively.

The *S* was told he would be shown a sheet con-

taining eight three-letter initials in specific positions, and the task was to learn the initials and their positions. After a 30-second study period, the sheet was removed and *S* was given a test page which was identical to the map, picture, or label sheet just studied except that the initials were omitted. The *S* had one minute in which to recall and write on the test page all of the initials, in their appropriate positions. After this test, a second 30-second study period and 1-minute test period were given, and so on until *S* was able to write all of the initials correctly, in their correct positions, on a single test. Thus in Stage 1 all *S*s in all conditions learned the initials and in addition *S*s in the Picture and Cognitive Groups had opportunity to associate these initials, respectively, with separated pictures or with parts of an integrated and potentially meaningful whole.

Forgetting interval. Following a perfect test trial in Stage 1, all *S*s were given a 10-minute forgetting interval which was filled by attempting to solve a letter puzzle. The puzzle consisted of a 12 × 12-cell matrix, with the letters A–L inserted in the 12 cells of the left column. The *S* was instructed to try to fill in the rest of the matrix cells with the letters A–L in such a manner that (*a*) no letter appeared twice in the same row, (*b*) no letter appeared twice in the same column, and (*c*) no letter appeared immediately before or after another letter more than once in the entire matrix.

Stage 2. Following the forgetting interval *S* was seated before a Stowe memory drum set in a black wooden screen, and was given instructions for the Stage 2 learning task. The learning task consisted of eight simple sentences, each beginning with a CVC which had been learned in Stage 1 and ending with a word indicating an occupation. The sentences for the CR conditions were: KOT is a pilot, NEB is a gas station attendant, BAF is a shopkeeper, PUM is a truckdriver, DOS is a policeman, RAZ is a cook, FER is a brakeman, and TUK is a farmer. For the OT conditions, the sentences were: KOT is a shepherd, PUM is a gold miner, DOS is a lumberjack, RAZ is a tollkeeper, NEB is a tailor, FER is a wealthy duke, BAF is a farmer, TUK is a fisherman.

The sentences were presented for 10 trials by the study-recall method. On each trial, the sentences were first presented one at a time in the drum window for study. Following the study period, the initials were presented alone, one at a time and in a different order from that used in the study period, and *S* was instructed to recall and say aloud the occupation of each initial as it came into view. At the end of each trial there was a 6-second rest before starting the next trial. Three different orders of presentation were used on successive trials to prevent serial learning.

To determine possible effects of study-recall time upon learning under the various cognitive conditions, half of the *S*s in each condition received the study and recall presentations at a 2-second rate and half at a 3-second rate. Within each condition, *S*s of both sexes were distributed evenly over these two levels of presentation time.

RESULTS

Since both sexes and a range of age levels were represented in each of the experimental treatments, preliminary analyses were made to determine possible effects of these *S* variables upon the data. Mean ages for each condition ranged between 15.75 and 16.16 years, a difference which did not approach statistical significance. To evaluate possible sex differences, the mean number of correct responses over the 10 trials of Stage 2 learning was calculated for boys and girls in the Cognitive, Picture, and Label conditions. These means, pooled over presentation times and the CR-OT variable, were $\bar{B} = 66.00$ and $\bar{G} = 64.83$ for the Cognitive condition, $\bar{B} = 60.25$ and $\bar{G} = 58.17$ for the Picture condition, and $\bar{B} = 61.25$ and $\bar{G} = 55.25$ for the Label condition. Although boys demonstrated a slight superiority over girls in Stage 2 performance, the highest *t* value obtained in comparing the means for each condition was 1.03 ($p > .10$). Therefore, the data from boys and girls was pooled in the subsequent analyses.

Stage 1 learning. To determine if the various presentation modes had differential effects upon Stage 1 learning, the mean number of Stage 1 trials taken to learn all stimuli was calculated for the OT-Cognitive, CR-Cognitive, OT-Picture, CR-Picture, OT-Label, and CR-Label groups. The means ranged from 2.33 to 2.92 trials, and did not differ significantly from each other, $F < 1.00$, $df = 5/66$. These data indicate that the Stage 1 learning task was relatively easy, and that neither the varying structure treatments nor differences in task stimuli (CR vs. OT) produced differences in speed of learning the initials.

Stage 2 learning. Table 1 presents means and *SD*s of number of correct responses over the 10 trials of Stage 2 learning for all groups. A preliminary test for homogeneity of variance indicated that differences in variance were not significant, $F_{max} = 17.47$, $df = 12/5$, $p > .05$. In a 3 × 2 × 2 analysis of variance, no significant difference was found in comparison of the

TABLE 1
MEANS AND STANDARD DEVIATIONS OF TOTAL CORRECT RESPONSES OVER 10 TRIALS OF STAGE 2 LEARNING

Groups	2 seconds				3 seconds			
	OT		CR		OT		CR	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Cognitive	61.00	3.90	66.50	10.95	68.33	4.89	67.50	8.36
Picture	51.83	12.48	50.33	16.00	63.17	10.38	71.50	5.96
Label	49.00	16.30	53.17	14.95	62.33	10.76	68.50	7.06

OT and CR conditions, $F = 1.99$, $df = 1/60$, $p > .05$. The difference between the 2-second and 3-second presentation rates was highly significant, as expected, $F = 20.15$, $df = 1/60$, $p < .001$. Also, as hypothesized, a significant difference was obtained for the main effect of cognitive structure, $F = 3.41$, $df = 2/60$, $p < .05$. The F value for the Presentation Time × Structure interaction was 2.11, $p > .05$, and F's of all other interactions were less than 1.00

Inspection of the individual group means in Table 1 shows clearly that performance of the Cognitive condition was superior to the performance means of the Picture and Label conditions at the 2-second presentation rate. At the 3-second presentation rate, however, the superiority of the Cognitive Group is less evident, and in fact is reversed among those groups receiving the CR task. The performance of the Cognitive groups was consistently high regardless of the rate at which the Stage 2 task was presented, whereas the Picture and Label groups learning the Stage 2 task at the 3-second rate were distinctly superior to their counterparts who learned at the 2-second rate. The consistently high performance of all groups at the 3-second rate suggests that the Stage 2 learning task was relatively easy and was learned to a near-ceiling level at the slower rate regardless of Stage 1 treatment, thus masking differences which showed up clearly when the task was made more difficult by decreasing the total learning time.

In view of these subtle differential effects of time upon performance, further statistical analyses were made to compare the effects of structure separately at each presentation rate. Since no significant difference between the CR and OT tasks was found in the first analysis, these tasks were pooled for the three structure conditions at each rate. Variation due to the simple main effect of structure at each level of rate was then evaluated, using the error term from the original analysis of variance according to a procedure described by Winer (1962, pp. 256–257). The results indicated a highly significant difference among the Cognitive, Picture, and Label conditions at the 2-second level of presentation, $F = 5.35$, $df = 2/60$, $p < .01$, but no significant difference among the structure conditions at the 3-second level, $F < 1.00$, $df = 2/60$. Subsequent individual comparisons among the structure conditions at the 2-second rate, using the Neuman-Keuls procedure (Winer, 1962, p. 238), demonstrated that the mean for the Cognitive condition was significantly higher than both the Picture and Label group means, $p < .05$, but that the latter two means did not differ reliably.

DISCUSSION

The performance of the Cognitive condition on the Stage 2 task was superior to both the Picture and Label conditions under all treatments except CR-3-seconds, and the differences obtained were statistically significant both in an overall test and at the 2-second level of presentation rate. These results, obtained with varying materials and a 10-minute rest interval between the Stage 1 and Stage 2 tasks, confirm and extend those of the previous study (Reynolds, 1966), suggesting that

the assumed cognitive structure imposed by prefamiliarization with a meaningful structure has relatively stable and general positive effects upon later learning. The finding of a general superiority of the Cognitive treatment over the Picture treatment also confirms the main hypothesis that it is the wholeness or integration of the total verbal-perceptual structure, and not simply the presence of associated verbal and pictorial material regardless of structure, which facilitates transfer to a related learning task.

The failure to obtain a statistically significant superiority of the Cognitive treatment at the 3-second rate of presentation requires further explanation, since it appears to detract from the clarity of the results. Table 1 shows that at the 3-second rate, all groups achieved means over 62, and four of the means were over 67. Since the highest possible score on the learning task was 80 correct responses over the 10 trials, it seems reasonable to consider that all of the groups receiving the slower rate were performing at or near the ceiling level for the task, and consequently significant differences among treatments were not observable at this rate. Alternatively, the 2-second rate made the learning task more difficult, allowing differences among the structure treatments to become evident. The data indicate that under this more difficult learning condition the performances of the Picture and Label groups drop considerably, while those of the Cognitive groups are maintained at near-ceiling level.

Thus it appears that the failure to obtain significant results at the 3-second rate was due to the ease with which materials were learned at that rate, and does not constitute a contradiction of the original hypothesis that prefamiliarization with an integrated and meaningful structure establishes a cognitive representation which will transfer positively to a rote learning task. Even so, further research using different Stage 1 conditions and more difficult Stage 2 learning tasks is desirable before a full specification of the effects of cognitive structure upon rote learning can be made.

REFERENCES

GLAZE, J. The association value of nonsense syllables. *Journal of Genetic Psychology,* 1928, **35,** 255–269.

PETERSON, L. R., & PETERSON, M. J. Short-term retention of individual verbal items. *Journal of Experimental Psychology,* 1959, **58,** 193–198.

REYNOLDS, J. H. Cognitive transfer in verbal learning. *Journal of Educational Psychology,* 1966, **57,** 382–388.

TOLMAN, E. C. Cognitive maps in rats and men. *Psychological Review,* 1948, **55,** 189–208.

WINER, B. J. *Statistical principles in experimental design.* New York: McGraw-Hill, 1962.

(Received April 14, 1967)

Education is a happening and here's what's happening in education

EDUCATIONAL PSYCHOLOGY
A Cognitive View

By DAVID P. AUSUBEL, The Ontario Institute for Studies in Education and the University of Toronto

The basic premise of this book is that educational psychology is primarily concerned with the nature, conditions, outcomes, and evaluation of classroom learning.

May 1968 **640 pp.** **$8.95 (tent.)**

CHILDREN
Readings in Behavior and Development

Edited by ELLIS D. EVANS, University of Washington

Designed to accompany *Children: Behavior and Development, Second Edition* by Boyd R. McCandless (Holt, Rinehart and Winston, 1967), this readings book is keyed to the chapter-by-chapter organization of that text, but also contains many readings which take up topics not mentioned in the text. In selecting the readings, the editor has emphasized research reviews, analytical papers, and research papers carrying implications for childrearing and childhood education.

April 1968 **576 pp.** **$5.95 paper**

LEARNING, LANGUAGE, AND COGNITION
Theory, Research, and Method for the Study of Human Behavior and Its Development

By ARTHUR W. STAATS, University of Hawaii

In this, the most comprehensive account available of learning and its relationship to the analysis of language, Professor Staats demonstrates that the principles and methods of experimental psychology—especially those of the psychology of learning—are also the building blocks from which to construct a general concept of human behavior.

January 1968 **640 pp.** **$9.50**

THE PSYCHOLOGY OF HUMAN GROWTH AND DEVELOPMENT
Second Edition

By WARREN R. BALLER, United States International University,
and DON C. CHARLES, Iowa State University

Updated and rewritten to meet the needs of today's students, the second edition of this popular book features a new chapter on thinking which covers the work of Jean Piaget and his Geneva school, and summarizes the principal contributions of American and British researchers.

March 1968 **480 pp.** **$8.95**

Holt, Rinehart and Winston, Inc.
383 Madison Avenue, New York, New York 10017

Important Books for Educators

Fp *An original paperbound series . . .*

CURRENT ISSUES AND RESEARCH IN EDUCATION

General Editor: Harry L. Miller, *Hunter College*

THE PSYCHOLOGY OF EDUCATION
Current Issues and Research

Edited by Donald H. Clark, *Hunter College*

Contains 52 selections concerning the culturally deprived, intelligence testing, creativity, emotional growth, ability grouping, the psychology of reading, national education assessment, and the classroom climate.

1967 288 pages paper, $2.95

ELEMENTARY EDUCATION
Current Issues and Research

Edited by Maurie Hillson, *Rutgers University*

Depicts the changing character of elementary education and includes discussions of emerging trends in foreign-language teaching, basic reading materials, mathematics, technology and computerized instruction.

1967 330 pages paper, $2.95

EDUCATION FOR THE DISADVANTAGED
Current Issues and Research

Edited by Harry L. Miller, *Hunter College*

Forty-nine selections on the world of the socially disadvantaged child stress urban and "inner city" problems in the schools and the increasing use of the behavioral sciences in dealing with these problems.

1967 288 pages paper, $2.95

SOCIAL FOUNDATIONS OF EDUCATION
Current Issues and Research

Edited by Dorothy Westby-Gibson, *San Francisco State College*

The selections in this volume examine the social role of education, pressures in childhood from kindergarten through college, the problems of socially disadvantaged youth, and the changing status of educators.

1967 313 pages paper, $2.95

SCHOOL CHILDREN IN THE URBAN SLUM
Readings in Social Science Research

Edited by Joan I. Roberts, *Hunter College*

The articles and excerpts present significant findings from anthropology, psychology, and sociology concerning problems in urban schools. Subjects given special attention include ethnicity, race, and socioeconomic factors in terms of how they affect intellectual potential, learning capacity, motivation, self-concept, and personality.

1967 640 pages $7.50

from The Free Press

THE COUNSELING OF COLLEGE STUDENTS
Function, Practice, and Technique

Edited by Max Siegel, *Brooklyn College*

Foreword by Harry D. Gideonse

The Counseling of College Students approaches college counseling from a practical point of view, concerning itself with the everyday realities of the counselor's world. With original contributions by 18 authorities in the field, this volume covers every significant aspect of student counseling.

January, 1968 487 pages $9.95

POLICY ISSUES IN URBAN EDUCATION

Edited by Marjorie B. Smiley and Harry L. Miller, both of *Hunter College*

The editors have chosen 28 readings for their relevance to policy formation on such issues as what should be taught in the "inner city" school, what changes must be made in teaching, and the need to remedy racial imbalance in urban schools. Among the authors represented are Frank Riessman, James S. Coleman, Irwin Katz, Whitney M. Young, Jr., and Kyle Haselden.

April, 1968 512 pages paper, $3.50

EDUCATION IN THE METROPOLIS

Edited by Harry L. Miller and Marjorie B. Smiley, both of *Hunter College*

Depicts the background of the disadvantaged urban child, confronting the reader with the human conditions and social implications necessary to an understanding of urban school problems. In addition, the book contains twelve pages of photographs which illustrate aspects of the environment described.

1967 303 pages paper, $2.95

TEACHING THE TROUBLED CHILD

By George T. Donahue and Sol Nichtern

"An interesting and well-written book in which [the authors] explain how 'teacher-moms' can help seriously disturbed school-age children . . . a challenging and provocative approach."—*Childhood Education. A Free Press Paperback.*

222 pages $2.45

THE FREE PRESS

A Division of The Macmillan Company

866 Third Avenue, New York, N. Y. 10022

WHAT IS DIFFERENT?

JOHN P. DE CECCO'S LEARNING AND INSTRUCTION: THE PSYCHOLOGY OF EDUCATIONAL PSYCHOLOGY

Dr. DeCecco of San Francisco State College, has carefully designed this book, with its systematic interrelationship of topics around a model of teaching, to be a maximally efficient and effective teaching instrument in its own right.

To make the study of the book profitable to students, the author provides lists of objectives at the beginning of each chapter, chapter section reminders of the points students should review, and questions within the text with fully explained answers.

Following closely Robert Gagné's classification of learning conditions, the author views the teacher as an intervener and arranger of appropriate learning conditions. The book gives more than traditional coverage to—theories and models of teaching, the linguistic and cognitive development of the average and disadvantaged child, the motivational functions of the teacher, the psychology of school learning as the basis for planning instruction, linguistics and psycholinguistics, all the current educational innovations in educational technology and the new science and mathematics, the construction and use of classroom and standardized test scores, and the evaluation of reports on educational research.

May 1968, approx. 800 pp., $9.75

For approval copies, write: Box 903 Prentice-Hall, Englewood Cliffs, N. J. 07632

THAT'S WHAT.

PRENTICE-HALL

WHY A SIXTH EDITION?

ARTHUR T. JERSILD'S CHILD PSYCHOLOGY

As a leader in its field for over thirty years, this book has been hailed in the United States and in translation in many European, South American, Middle and Far Eastern countries.

But now Arthur T. Jersild of Columbia University, has added entirely new chapters or sections on genetic factors, prenatal development, the influence of stimulation and deprivation in infancy, the origins of intelligence, past and recent, and research dealing with cognitive development, particularly inspired by Piaget.

Other chapters such as those on the self, the study of dreams, anxiety, language development and the study of perception are substantially rewritten or revised. The book remains sensitive and readable. There is a continued emphasis on the self as essential in understanding all aspects of developmental psychology.

And, there is a distinctly human touch.
February 1968, 640 pp., $8.50

THAT'S WHY.

For approval copies, write: Box 903
Prentice-Hall, Englewood Cliffs, N. J.
07632

PRENTICE-HALL

THE PSYCHOLOGICAL FOUNDATIONS OF EDUCATION SERIES

General Editor: Victor H. Noll, Michigan State University

Intensive investigation in the field of educational psychology has resulted in the establishment of a wide variety of significant topics that warrant independent consideration. This paperbound series is designed to provide a deeper examination of specific areas which are not given thorough treatment in many comprehensive textbooks. The volumes in the Noll series offer a convenient source of up-to-date and expertly written supplementary material. Teachers and students will benefit from the advanced coverage of pertinent subject matter. Used in conjunction with a comprehensive text, these books will increase the impact and effectiveness of current courses in the psychological foundations sequence. Theoretical material is an important part of each book, but the emphasis is centered on practical applications.

Psychology of Adolescence for Teachers
By Glenn Myers Blair and R. Stewart Jones, both, University of Illinois
1964, 128 pages, $1.50

Psychology of the Child in the Classroom
By Don C. Charles, Iowa State University
1964, 86 pages, $1.25

The Psychology of Learning in the Classroom
By Robert C. Craig, Michigan State University
1966, 85 pages, $1.25

The Mentally Retarded Child in the Classroom
By Marion J. Erickson, Ypsilanti Public Schools
1965, 114 pages, $1.25

The Psychology of Discipline in the Classroom
By William J. Gnagy, Illinois State University
1968, 80 pages, $1.50

Problem Solving in the Classroom
By Bryce B. Hudgins, Washington University
1966, 74 pages, $1.25

Teacher Self-Evaluation
By Ray H. Simpson, University of Illinois
1966, 100 pages, $1.25

Guidance in the Classroom
By Ruth Strang, University of Arizona, and Glyn Morris, Asst. Supervisor of Guidance and Curriculum, Lewis County, N. Y.
1964, 118 pages, $1.50

Gifted Children in the Classroom
By E. Paul Torrance, University of Minnesota
1965, 102 pages, $1.25

Write to the Faculty Service Desk for examination copies:

THE MACMILLAN COMPANY 866 Third Avenue, New York, New York 10022
In Canada, write to Collier Macmillan, 55 York Street, Toronto 1, Ontario

Five from Macmillan

General Psychology
By David C. Edwards, Iowa State University

"Long overdue . . . a one term book such as this cannot miss. I find the book to be very well written, direct, and to the point."—James Bruning, Ohio University
This is a brief yet thorough introduction to psychology—its content, its terminology, and its methods. The author has accomplished brevity by carefully selecting the important topics, and by avoiding unnecessary or irrelevant exposition and discussion of the obvious. The approach is modern, and is concerned with psychology as a science of behavior. The text is suitable for either a brief or a standard course in general psychology in which supplementary materials are used.

1968, 384 pages, $5.95

Statistics in Education and Psychology: A First Course
By Merle W. Tate, Lehigh University

Basic instruction is here provided in most of the techniques used currently in measurement and research, with emphasis on critical analysis, proper use of statistics, and the intelligent interpretation of results. Most of the topics are presented through examples which are completely developed and interpreted through the use of actual data. Sampling notions are introduced early and treated more rigorously in final chapters.

1965, 352 pages, $5.95

Statistics for the Classroom Teacher: A Self-Teaching Unit
By Edward Arthur Townsend, Westfield State College, and Paul J. Burke, The City College of The City University of New York

This supplementary text provides a review of the basic fundamentals of statistical concepts and procedures. The text explains the basic procedures for collecting and organizing data of the type most frequently encountered by the classroom teacher. A simple explanation of each statistical concept and process is followed by exercises for self-rehearsal and self-teaching, thereby saving valuable class time.

1963, 80 pages, paper, $1.95

The Essentials of Learning, An Overview for Students of Education Second Edition
By Robert M. W. Travers, Western Michigan University

Thoroughly updated, this edition provides unusually broad coverage of the field of learning, and discusses individual topics within the field in specific detail. Thus, not one theory of learning is presented, but many, each firmly rooted in research, giving the student significant and current data on every aspect of learning.

1967, 544 pages, $7.95

An Introduction to Educational Research Second Edition
By Robert M. W. Travers

In this thoroughly revised edition of a classic text, Professor Travers provides an interpretation of the aims and methods of educational research. The role of theory is emphasized throughout and distinguishes this text from others in the field.

1964, 608 pages, $7.50

Write to the Faculty Service Desk for examination copies

THE MACMILLAN COMPANY 866 Third Avenue, New York, New York 10022
In Canada, write to Collier Macmillan, 55 York Street, Toronto 1, Ontario

Educational Psychology

SECOND EDITION

by *LEE J. CRONBACH, Stanford University*

A leading textbook for educational psychology courses, marked by pronounced emphasis on intellectual learning and development and a full presentation of the psychological theory and research that underlie educational practice. Dr. Cronbach focuses on the necessity for understanding the pupil as a person, showing how the most useful learning theory is one that stresses the learner's purposes and interpretations. Accompanied by a Student Guide and two separate Test Item Files.

706 pages. $8.95

A Social Psychological View of Education

by *CARL W. BACKMAN* and *PAUL F. SECORD, University of Nevada*

A clear and thoroughly documented review of the concepts and data from social psychology which are directly pertinent to the role of the teacher in the classroom. The authors consider the teacher, the student, and the school as part of a social system, emphasizing how important social experience is to the educational process.

Paperbound. 125 pages. $2.25 (probable)
Publication: May 1968

Learning

by *J. CHARLES JONES, Bucknell University*

This concise interpretation of recent psychological research findings on learning provides the teacher-in-training with a theoretical basis for examining procedures and problems of education.

Paperbound. 179 pages. $2.25

Measuring Pupil Achievement and Aptitude

by *C. M. LINDVALL, University of Pittsburgh*

This volume provides an introduction to the basic principles of testing and evaluation that are essential to teachers for the effective assessment of pupil achievement and aptitude.

Paperbound. 188 pages. $2.45

For more detailed information, please write the publisher.

HARCOURT, BRACE & WORLD, INC.
New York / Chicago / San Francisco / Atlanta

Consulting Psychologists Press presents a full range of techniques for appraising developmental level of pre-school, nursery, kindergarten and early elementary children.

A parent-administered device —

• *THE SCHOOL READINESS SURVEY*

An ingenious new way to involve parents in objectively appraising their own children's readiness and in expediting their children's mental growth.

Contains 7 easily administered and scored subtests to survey skills required in school, plus dozens of suggestions to parents for helping their children develop. Parent administration saves staff time and increases parental understanding of their children and the learning process.

Send $1 for Specimen Set

A teacher-administered survey —

• *VALETT DEVELOPMENTAL SURVEY OF BASIC LEARNING ABILITIES*

Contains 233 easily administered tasks from which to choose, covering 7 areas: Motor Integration & Physical Development, Tactile & Auditory Discrimination, Visual-Motor Coordination, Visual Discrimination, Language Development & Verbal Fluency, and Concept Development. For teacher use with children aged 2 to 7.

Send $1 for Specimen Set (Demonstration Cards $3 extra)

For the psychologist or specially trained teacher —

• *MARIANNE FROSTIG DEVELOPMENTAL TEST OF VISUAL PERCEPTION*

For more refined evaluations of perceptual skills, this widely-used test yields scaled scores in 5 areas for children aged 4 to 8.

Send $5 for Complete Specimen Set

We distribute The Frostig Program for Development of Visual Perception and the new Pictures and Patterns.

For the school psychologist —

• *PSYCHOEDUCATIONAL PROFILE OF BASIC LEARNING ABILITIES*

Not a test but an 8-page profile for summarizing clinical information and test data in 5 basic ability areas. Aids the psychologist in discussing his findings with parents and teachers.

Send .75 for Complete Specimen Set

For evaluating social competency of the mentally retarded —

CAIN-LEVINE SOCIAL COMPETENCY SCALES

A 44-item scale to estimate social competence in the trainable mentally retarded, chronological age 5–13. Subscales for Self-Help, Initiative, Social Skills, and Communication. A useful aid in diagnosis, placement, and training.

Send $1.50 for Specimen Set

CONSULTING PSYCHOLOGISTS PRESS

577-D College Avenue, Palo Alto, California 94306

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

David A. Payne and Robert F. McMorris

For courses focusing on tests and measurements, this text contains fifty-four theoretical and research papers providing a broad perspective on test development with an emphasis on the assessment and prediction of learning outcomes. A manual of examination questions is available. 1967 419 pages Paper $5.75

THE SPECIFICATION AND MEASUREMENT OF LEARNING OUTCOMES

David A. Payne

Intended for undergraduate and graduate courses, the text has a single purpose: to provide the classroom teacher with a practical and efficient set of techniques to aid in evaluating student achievement. Essential principles of educational and psychological measurement are presented along with many concrete examples. Abbreviated but valid statistical techniques are described with emphasis on their application for improving test analysis and interpretation. 1968 209 pages Paper $2.75

STUDIES IN EDUCATIONAL PSYCHOLOGY

Raymond G. Kuhlen

For courses focusing on the psychology of school learning, this text contains fifty papers dealing with the learning process with special attention to the cognitive, personality, and motivational factors involved. 1968 482 pages Paper $5.75

READINGS IN THE PSYCHOLOGY OF CHILDHOOD AND ADOLESCENCE

William J. Meyer

For courses focusing on human development, this book contains sixty-five theoretical papers and research reports dealing with the main features of child and adolescent development. 1967 436 pages Paper $5.25

THE PSYCHOLOGY OF VOCATIONAL CHOICE

John L. Holland

This text offers a concise synthesis of current knowledge of vocational adjustment, and further presents original theory and suggestions for new research. Written primarily for students and professional audiences, the book effectively interprets data for practical application. 1966 132 pages Paper $1.95

METHODS OF STUDYING THE INDIVIDUAL CHILD: THE PSYCHOLOGICAL CASE STUDY

John Rothney

Intended for teachers, counselors and school psychologists in training and in service, this text advocates the case study as a primary aid in understanding and furthering the development of the individual child. The individuality of each child is stressed as is the importance of using a wide variety of techniques to present, in case study form, as accurate a picture of the child as possible. Fall 1968

BLAISDELL PUBLISHING COMPANY

A Division of Ginn and Company

275 Wyman Street, Waltham, Massachusetts 02154

Chandler Publications in the fields of education and psychology

NEW **Discipline in the Classroom: Basic Principles and Problems**
by Staten W. Webster,
University of California, Berkeley
The first part of the book examines the causes of student malbehavior. Pointed out to the reader is the necessity to view acts of malbehavior as products of possible interactions among several factors including the personality of the student, the personality of the teacher, and the particular human physical environment of the classroom.
The second part of the book offers case reports of student behavior problems. The reports include analyses of the problems and proposed suggestions for solving the problems written by highly regarded and experienced teachers.
Approximately 148 pages 5½″ x 8¼″
© 1968 paper List $2.70

Instruction: Some Contemporary Viewpoints
edited by Laurence Siegel,
Louisiana State University
The reader is familiarized with a substantial body of knowledge, provided with organizing concepts and principles, and is exposed to the most recent work and thinking that has been done in the field of education.
390 pages 5½″ x 8¼″ © 1967 cloth
List $7.00

Motivation: Psychological Principles and Educational Implications
by Melvin H. Marx, University of Missouri
and Tom N. Tombaugh, Carleton University
The book outlines the major efforts made thus far by psychologists to explain the multifold facts of what is commonly called "motivation". The text provides students with a scientific explanation of the origin of such terms as ego, drive and aggression, while examining the empirical and theoretical methodology of the professional psychologist.
304 pages 5½″ x 8¼″ © 1967 paper
List $2.95

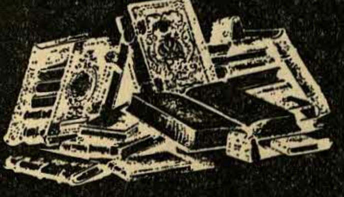
Learning: A Survey of Psychological Interpretations
by Winfred F. Hill, Northwestern University
The emphasis in this book is on learning theory to provide basic understanding of the learning process. Psychological interpretations of learning are classified into the familiar divisions of connectionist and cognitive and some of the attempts that have been made to combine the advantages of these two categories.
238 pages 5½″ x 8¼″ © 1963 paper
List $2.50

From the Chandler Publishing Company,
San Francisco
Distributed by

SRA

Science Research Associates, Inc.
259 East Erie Street, Chicago, Illinois 60611

A Subsidiary of IBM



THE ONTARIO INSTITUTE FOR STUDIES IN EDUCATION

102 Bloor Street West • Toronto 5 • Ontario • Canada

Recent publications from the Department of Applied Psychology

Recent Research on the Acquisition of Conservation of Substance
Educational Research Series No. 26, 1967, pp. v + 72, $1.75
Edited by David W. Brison and Edmund V. Sullivan. Three investigations.

Learning Theory and Classroom Practice
Bulletin No. 1, 1967, pp. v + 34, $1.75
By David P. Ausubel. Examines current learning theories.

Piaget and the School Curriculum: A Critical Appraisal
Bulletin No. 2, 1967, pp. vii + 38, $1.25
By Edmund V. Sullivan. Implications of Piaget's theory of intellectual development.

Accelerated Learning and Fostering Creativity
1968, pp. iii + 40, $1.50

Edited by David W. Brison. Four papers from the Phi Delta Kappa—OISE Centennial Symposium, Toronto, February 1967.

Write PUBLICATIONS, OISE, 102 Bloor St. W., Toronto 5, Canada

McGraw-Hill Books

CLASSICAL PSYCHO-PHYSICS AND SCALING

By Sidney A. Manning, University of Maryland; and Edward Rosenstock, Pennsylvania State University.

176 pages, $2.50

This is a semi-programmed text designed to teach psychophysics and scaling as part of a general introductory course, or as an adjunct to courses in Perception, Experimental, or Physiological Psychology.

READINGS IN PSYCHO-LOGICAL FOUNDATIONS OF EDUCATION

By Walter H. MacGinitie and Samuel Ball, Teachers College, Columbia University

Available Spring, 1968

The authors have selected articles that are understandable to college students with little or no background in psychology or statistics. The book is especially effective for use in a one-semester course.

CLINICAL TEACHING

By Robert M. Smith, Pennsylvania State University

Available Spring, 1968

In this text, the author considers the educational problems of the retarded child from the pre-school through the post-school periods. Suggestions are made for working with parents of the retarded within the structure of the school program.

Examination copies upon request

McGraw-Hill Book Company

**330 West 42nd Street,
New York, New York 10036**

new from bobbs-merrill

THE PREDICTION OF ACHIEVEMENT AND CREATIVITY

by RAYMOND B. CATTELL and H. J. BUTCHER

The authors have developed and utilized empirical methods to observe and analyze basic personality factors. In this new work they survey the present state of knowledge in the field and report findings of their research on the personality and abilities of school children. In addition, they outline a specific program for testing students at key points during the school years, and offer suggestions for stimulating greater creativity and achievement within an educational system.

"...this is a top-quality piece of work and reflects highly creative and original thinking on the part of the authors"

—John D. Krumboltz,
Stanford University

400 pages — $7.50 clothbound, on approva

previously published

THE PSYCHOLOGY OF LEARNING APPLIED TO TEACHING

by RICHARD B. BUGELSKI

"...solid, accurate, wise and intelligent. The summary of knowledge about the field of learning is obviously based on thorough scholarship."

—Thomas K. Landauer,
Stanford University

302 pages — $6.00 clothbound

Accompanying manual with examination questions available upon request

THE PSYCHOLOGY OF ADOLESCENCE

by MARVIN POWELL

"...a very superior book...one of the best written, most interesting, and most readable that I have seen."

—Glenn M. Blair,
University of Illinois

576 pages — $6.50 clothbound

college division

THE BOBBS-MERRILL COMPANY, INC.
A SUBSIDIARY OF HOWARD W. SAMS & CO., INC.
4300 WEST 62nd STREET • INDIANAPOLIS, INDIANA 46268

JOURNAL OF
EDUCATIONAL PSYCHOLOGY

Volume 59 | June 1968 | Number 3

CONTENTS

© 1968 by the American Psychological Association, Inc.

Journal of Educational Psychology
1968, Vol. 59, No. 3, 139-146

LIBRARY

CERTAIN EFFECTS OF THE EXPECTATION TO TRANSMIT ON CONCEPT ATTAINMENT[1]

PAUL DAVIDSON REYNOLDS[2]
Stanford University

36 undergraduate women studied analytical concepts under 3 conditions: (a) In the Alone condition (AC), 10 Ss studied the concepts expecting to be tested on their knowledge of them. (b) In the Peer condition (PC), 13 Ss studied the concepts expecting to transmit them to an undergraduate woman. (c) In the Child condition (CC), 13 Ss studied the concepts expecting to transmit them to a 6-yr.-old boy. All Ss, the "transmitters," after one-way verbal communication to their receivers, were tested on their knowledge of the concepts. The Ss were divided into exceptionally high (EHM) and above average math (AAM) aptitude groups on the basis of their math aptitude scores on the College Entrance Examination Board tests. The performance scores of the AAM Ss in the PC and CC were lower ($p <$ .05) than the scores of the AAM *S*s in the AC.

Recent summaries of research on cognitive processes, problem solving, and thinking (Bruner, Goodnow, & Austin, 1962; Duncan, 1959; Gagné, 1959; Kendler, 1961; Leeper, 1951) are notable in that they reflect an absence of attention to the subject's (*S*'s) purpose in acquiring concepts, whether the individual is learning for his own purposes or expects to relay his newly acquired knowledge to another. Zajonc (1960) and Cohen (1961), interested in examining the nature of cognitive structures that are activated or "tuned in" when persons enter into communication with others, have studied the effect of expecting to transmit information compared with the effect of expecting to receive additional information on the way an individual organizes information he has received. The purpose of this study is to consider if the expectation to transmit knowledge will reduce the capacity of an individual to acquire difficult-to-describe material.

Given the basic situation of an *S* learning a concept and then transmitting it to another, the following is one possible conceptualization of the individual's cognitive processes. Assume that the individual has two types of closely related repertoires: (*a*) A *conceptual repertoire* of approaches, strategies, logical systems, or methods of analysis that he utilizes when attempting to solve a problem or understand any type of subject matter. (*b*) A *symbolic repertoire* of signs, words, symbols, and phrases that he uses to code the concepts in the conceptual repertoire, to store information of any type, and to transmit concepts to another individual.

This conception of a symbolic repertoire implies that any mental activity of the individual, conscious or unconscious, utilizes symbols from the symbolic repertoire, particularly when he uses concepts from the conceptual repertoire. Considering only those symbols the individual uses in transmitting information and ideas to another individual, the following can be defined: The *transmission vocabulary* consists of all those symbols, words, signs, or phrases that the individual uses to communicate with or to transmit to another individual. An implication of this conceptualization is that the symbols that compose the transmission vocabulary are a subset of the symbolic repertoire.

[1] This research was originally done in partial fulfillment of the requirements for the master's degree in psychology. The author wishes to thank his advisor, Alex Bavelas, who provided the intellectual stimulation that made this investigation possible and maintained the guidelines that kept the development of these ideas oriented in a fruitful direction. Thanks are also due to Louise R. Pierce. She solicited most of the subjects from the student volunteer tutors who participated in the Stanford Tutorial Project, of which she was the director. Nanci Irene Moore was kind enough to provide some of the original data from her research for her master's thesis.

[2] Currently with the Department of Sociology, Stanford University.

Copyright © 1968 by the American Psychological Association, Inc.

The present experiment investigates the effect an expectation to transmit will have on the learning performance of an individual. For ease of analysis, two situations can be compared: (*a*) The learning-without-expectation-to-transmit situation, where the individual learns a subject matter with no expectation that he will transmit it to a receiver. (*b*) The learning-with-expectation-to-transmit situation, where the individual learns a subject matter expecting to transmit it to a receiver. Only when the transmitter has completed his learning does he transmit his newly acquired knowledge to a receiver.

Considering the two situations described above, there are two questions of interest: (*a*) Is the learning of the individual impaired in the learning-with-expectation-to-transmit situation as compared to the learning-without-expectation-to-transmit situation? (*b*) Is there a loss of knowledge in the transmission between the transmitter or "learner" and the receiver? Based on these questions, there are four possible cases of the effects of the learning-with-expectation-to-transmit situation when it is compared to the learning-without-expectation-to-transmit situation. These are arrived at by first considering that the learning of the subject may not be impaired (L) or may be impaired (−L) and then considering that there is no loss in the transmission of subject matter (T) or that there is a loss in transmission (−T).

The four possible cases of the learning-with-expectation-to-transmit situation are as follows:

1. In the L,T case there would be no reduction in *S*'s learning of the material when compared to the learning-without-expectation-to-transmit situation. After transmission to a receiver, the receiver demonstrates knowledge of the material equal to that of the transmitter.

2. In the L, −T case there would be no reduction in *S*'s learning of the material. After transmission to a receiver, the receiver demonstrates *less* knowledge of the material than the transmitter.

3. In the −L,T case there would be a reduction in *S*'s learning of the material when compared to the learning-without-expectation-to-transmit situation. After *S*'s transmission to a receiver, the receiver demonstrates knowledge of the material equal to that of the transmitter.

4. In the −L, −T case there would be a reduction in *S*'s learning of the material when compared to the learning-without-expectation-to-transmit situation. After *S*'s transmission to a receiver, the receiver demonstrates less knowledge of the material than the transmitter.

If the L,T case occurs, where there is no impairment of learning or transmission, it may be assumed that the expectation to transmit has no observable effect on the individual's cognitive processes.

The L, −T case, where there is no impairment of learning but there is impairment of transmission, would imply that the individual used his conceptual repertoire with facility in learning the subject matter, but that some of the symbols used in learning the subject matter were not a part of the transmission vocabulary. In other words, the individual was readily able to learn concepts and ideas, but was not able to express ideas that he understood. This reasoning would imply that the symbolic repertoire includes more symbols than the transmission vocabulary, which is an issue of some interest.

If the −L,T case occurs, where there is an impairment of learning but no impairment in transmission, the reduction in learning performance may be attributable to either or both of two effects: First, if we assume that the transmission vocabulary is smaller than the total symbolic repertoire, it may be that *S*, knowing he will be required to transmit the soon-to-be-learned subject matter to another, will restrict his conceptual repertoire to those concepts that are represented by symbols in the transmission vocabulary. In other words, it may be that in attempting to understand the subject matter the individual utilizes only those concepts or schemata that he can transmit to the receiver with ease and confidence.

Second, it may also be that *S* is not sure that the receiver has a conceptual repertoire that is equivalent to his and may restrict his conceptual "tool kit" to only

those concepts that he is sure that the receiver will understand, transmission of the concepts being of minor concern. For example, the easiest way to represent algebraically a particular geometric figure may be to utilize polar coordinates. But if the transmitter perceives that the receiver is only accustomed to rectangular coordinates, he may attempt to utilize the inappropriate rectangular coordinates and as a result, may fail to understand the geometric figure himself.

If the −L, −T case occurs, where there is an impairment of learning and an impairment of transmission, all three of the effects described in the two previous cases (L, −T and −L, T) could be occurring.

Finlay (1966) studied the difference in the effect on a receiver's performance after he received written instructions from a transmitter who had just had an opportunity to learn a skill—arranging a 2-foot chain to achieve a high score. The major result was that if the transmitter and the receiver did *not* know in advance that the transmitter was to provide the written instructions, the receiver did significantly better on the task than if the transmitter and receiver did know in advance that the transmitter would provide written instructions for the receiver.

The results of the Finlay experiment allow speculation that, given certain types of subject matter, an individual in the learning-with-expectation-to-transmit situation will transmit less information to a receiver than an *S* in the learning-without-expectation-to-transmit situation. In the above discussion of this case (L, −T) we concluded that the occurrence of this phenomenon would imply that the transmission vocabulary was smaller than the symbolic repertoire.

The purpose of the present study is to determine if the other possible cases, −L, T or −L, −T, of the learning-with-expectation-to-transmit situation could occur for certain types of subject matter. Specifically, the interest is the possibility of a decrement in the learning performance of *S*s when they expect to transmit. The major question asked is: Is the test performance of *S*s after acquiring concepts affected when they expect in advance to transmit the concepts?

If there is a reduction in the amount of learning that an individual exhibits in the learning-with-expectation-to-transmit situation, it should be strongest when the characteristics of the receiver lead the individual to reduce either his transmission vocabulary or his conceptual repertoire. Therefore, an additional question will be addressed: Will the individual's perception of the ability of the receiver to understand his transmission have an effect on the individual's test performance after acquiring concepts?

THE CONCEPTUAL PROBLEM

The concepts were analytical in nature and are discussed in detail in Reynolds (1966, Appendix II). The *S*s attempted to determine the concepts common to each of four categories of designs in the learning set, Figure 1. The two images in each rectangle represent a single design. The colors (green, blue, red, and yellow) were used only for identification of the categories.

If an individual has an opportunity to learn a subject matter, and can demonstrate knowledge of the material through nonverbal responses, the translation of ideas or concepts into words, or symbols from the transmission vocabulary, is unnecessary. It may be possible to assume that the transmission vocabulary is being bypassed and that knowledge of the subject matter is being measured directly. For this reason, all *S*s demonstrated knowledge of the concepts by sorting designs, similar to those in the learning set, into four categories without access to the learning set.

METHOD AND PROCEDURE

The main factor that is expected to affect the learning performance of *S* is the expectation that what is learned be put into words. In addition, it is expected that the transmitter's perception of the size and sophistication of the receiver's conceptual repertoire and transmission vocabulary may have an influence on the magnitude of the effect. To investigate the second issue, an attempt was made to manipulate the transmitter's perception of the (*a*) experience and (*b*) development of the receiver's conceptual repertoire and transmission vocabulary.

For the Peer condition, it was assumed that *S*s

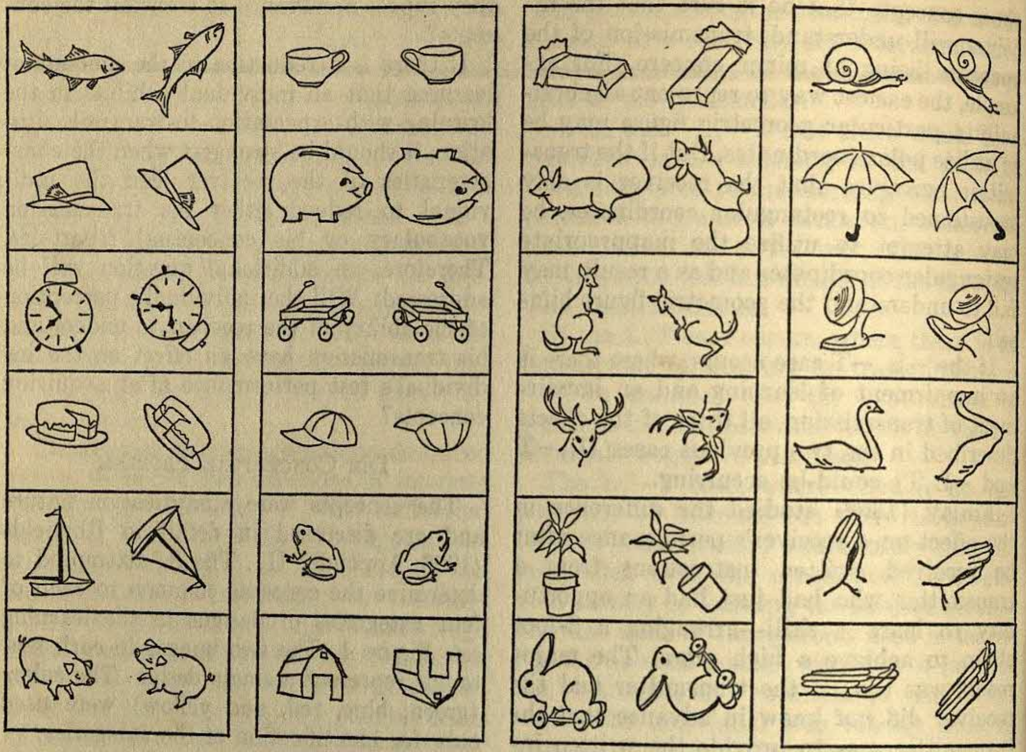


FIG. 1. Task learning set. (Columns from right to left are green, blue, red, yellow.)

would perceive another individual of like sex and approximately equal social and educational background as having an equally developed conceptual repertoire and transmission vocabulary. For the Child condition, it was assumed that *Ss* would perceive a child as having a less developed conceptual repertoire and transmission vocabulary.

Since it was impossible to test how much each *S* would have learned if they had not expected to transmit, an Alone condition was used to measure the learning performance of comparable *Ss* when they did not expect to transmit.

The experiment involved three conditions: The *Peer condition* (PC). The *S*, expecting to instruct a peer, studied the learning set, instructed the peer receiver, and without prior notice, was asked to take the test. The *Child condition* (CC). The *S*, expecting to instruct a 6-year-old boy, studied the learning set, instructed the child receiver, and without prior notice, was asked to take the test. The *Alone condition* (AC). The *S*, expecting to be tested herself, studied the learning set and then took a test of her knowledge when she indicated she was ready.

Thirteen *Ss* experienced the PC, another 13 *Ss* experienced the CC, and 10 *Ss* experienced the AC.

Subjects

The 49 undergraduate women that served as *Ss*, 36 as transmitters and 13 as peer receivers, were volunteers recruited in three ways: 32 from a group of unpaid tutors who were working with potential dropouts in public schools, 14 from undergraduate courses, and three were friends of *Ss* participating in the research.

It was decided that it was best to make the manipulation in the case of the child receivers as strong as possible without arousing the suspicion of *Ss*. Therefore, a first grade child was selected as the lowest level of perceived intellectual development that was high enough so that *Ss* might feel that they had some chance of explaining the concepts to a child receiver. Two child confederates were used, both 6-year-old boys. They received 25 cents for each participation in the experiment and were fully aware of the part they were expected to play.

Peer Condition

When the two women arrived for the PC, the roles of the transmitter and the receiver were randomly assigned and the receiver was asked to wait outside the room while the transmitter received the following instructions:

1. The transmitter was told that the purpose of the study was to determine how well the transmitter could transfer concepts to the receiver. The transmitter was told that she was the receiver's only source of information and that the major

measure of performance was the receiver's test score.

2. Using a sample task it was demonstrated that although the actual designs in the learning set and test set were different, they could be grouped on an abstract basis. All *S*s completed the sample task without errors.

3. The transmitters were told that the only communication allowable between them and the receiver would be their verbal instructions. No visual cues to nor any verbal utterances on the part of the receiver were allowed.

4. The time limits were described in detail. The transmitter was to be given a maximum of 35 minutes to study the concepts and instruct the receiver. This 35 minutes was divided into two parts, a study period which could last up to 20 minutes during which the transmitter was not required to say anything and an instruction period which lasted until the transmitter voluntarily terminated the instructions or 35 minutes had passed. The transmitter was told that the receiver would be given all the time she required to complete the test.

The complete set of instructions was always read to the transmitter (taking about 20 minutes) who was encouraged to ask questions about any part that she did not understand. The transmitter and receiver were permitted to ask questions about procedure during the experiment.

Following the instruction of the transmitter, the receiver was brought back into the room and seated at a table back-to-back with the transmitter. She was then given an abbreviated set of instructions which stated that the receiver was to separate cards in accordance with the instructions of the transmitter and that she was not allowed to say anything about the task or see what the transmitter was doing. The receiver was told that the focus of the study was on her test performance and that once she had received her test set, she would be allowed all the time she wanted to complete the test.

Upon completion of the instructions, the transmitter was immediately given the learning set, and the procedure described to the transmitter was followed until the transmitter's instructions to the receiver were completed. When the transmitter voluntarily terminated her instructions, or the 35 minutes had elapsed, the learning set was removed from the transmitter and the transmitter was given a test set and asked to separate the designs into the proper categories, taking as long as desired. At no time did any *S* appear to be hampered by the time limits.

The transmitter was encouraged to be as accurate as possible. When she had finished the test, she was asked a series of questions about the experiment and the entire purpose of the study was explained to both transmitter and receiver.

Child Condition

Because the child receivers were confederates, the design of the procedure was intended to minimize the contact between the transmitter and the child receiver until after the transmitter had taken the test. In the CC, the experimental procedure was exactly the same as that in the PC except that the child receivers made no attempt to take the test. Although no transmitters were experiencing close and continuous contact with children outside the experimental situation, none showed any apprehension or nervousness about teaching a child.

Alone Condition

In the AC, the procedure was varied by eliminating all instructions that pertained to the receiver, retaining the remainder of the instructions, including the sample problem. The time limits remained the same except that it was necessary to put a 10 minute minimum time limit on the study session, while retaining the 35 minutes maximum time limit. The *S*s in this condition knew from the start that they would be tested on their knowledge of the concepts.

The last question of the postexperimental interview was: "When did you first realize that you were going to take the test yourself?" For all transmitters in the PC and CC conditions, the answer was, in effect, "When you handed me the cards."

Scoring of Conceptual Attainment

Since the most important data collected are the scores that the transmitters achieved on the test of their knowledge of the concepts in the learning set, it is proper to consider what this score actually means in terms of conceptual acquisition. It seemed appropriate to design a procedure that would permit a count of the number of concepts acquired, utilizing this figure as an ordinal measure of performance. The following assumptions resulted in a procedure that allowed the translation of the numerical score, which could vary from 0 to 72, to a measure of the number of concepts acquired, which could vary from 1 to 4.

Assuming that an *S* had no knowledge of the concepts and distributed the 72 test designs to the four categories on a random basis, it would be expected that 18 of the designs would be correctly classified by chance ($72/4 = 18$), Level 1.

Assuming that an *S* understood one concept well, placed 18 designs into one correct category, and distributed the remaining 54 designs on a random basis; it would be expected that 18 designs would be correctly placed by chance into the remaining 3 categories for a total score of 36 (18 [from knowledge] + 54/3 [by chance] = 36), Level 2.

Assuming that an *S* understood two concepts well, placed 36 designs into two correct categories, and distributed the remaining 36 designs on a random basis; it would be expected that 18 designs would be correctly placed by chance into the remaining 2 categories for a total score of 54

(36 [from knowledge] + 36/2 [by chance] = 54), Level 3.

Assuming that an *S* understood three concepts well, placed 54 designs into three correct categories, and put the remaining 18 designs into one group; it would be expected that a score of 72 would result (54 [through knowledge] + 18 [by default] = 72). Finally, assuming that an *S* understood all four concepts, it would be expected that all 72 designs would be correctly placed into the four categories for a score of 72. Since it is impossible to separate these two levels of conceptual attainment through examination of the test score alone, both of these occurrences are considered Level 4.

The procedure used for making inferences about the number of concepts attained by *Ss* from their performance score was as follows: The *Ss* who scored between 0 and 27 were placed in Level 1. The *Ss* who scored between 28 and 45 were placed in Level 2. The *Ss* who scored between 46 and 63 were placed in Level 3. The *Ss* who scored between 64 and 72 were placed in Level 4. There was no problem in classifying scores with this procedure since, with one exception, *Ss'* scores were clustered around 36, 54, or 72.

After translating *Ss'* scores into levels of conceptual attainment, it is appropriate to consider this an ordinal measure of performance. Siegel (1956, p. 136) describes the Kolomogorov-Smirnov two-sample test and a chi-square approximation for small unequal sample sizes which appears to meet the assumptions of the ordinal measure of conceptual attainment.

RESULTS

Comparison of the transmitters' performance level scores across the three experimental conditions shows no statistically significant differences and no clear direction of improvement, although the best performances were in the Alone condition.

As the test concepts appeared to be analytical in nature, it seemed reasonable to expect that analytically oriented individuals might be more successful at understanding the test concepts. For an independent measure of the analytical aptitude of the transmitters, the mathematical aptitude scores on the College Entrance Examination Board tests were obtained. Since the resulting distribution of math aptitude scores for the transmitters was bimodal and showed a break of 13 points (635–648) at the average score of 648, the transmitters were divided into an "exceptionally high math" (EHM) group of 20 *Ss* with math aptitude scores of 650 to 770 and an "above average math" (AAM) group of 16 *Ss* with scores ranging downward from 650.

The performance scores of the transmitters, after they had been classified into the two groups, is shown in Table 1. The change in performance among the AAM transmitters is statistically interesting and indicates that the expectation to transmit may have an effect on the learning performance of the transmitter, for the AAM transmitters in both the PC and CC have lower performance scores than the AAM transmitters in the AC.

If the same procedure is followed using the verbal aptitude scores of the transmitters, the results among the six cells shows the same trends as found with the separation based on the math aptitude scores. But the statistical significances are much lower. This similarity may be due in part to the high correlation between the math and verbal aptitude scores. Using a simple linear regression to obtain a Pearson *r* and utilizing 50 undergraduate women on whom data

TABLE 1

NUMBER OF TRANSMITTERS ATTAINING EACH PERFORMANCE LEVEL

Experimental condition	Performance level attained	Transmitters' Math Aptitude: Above average (AAM) Number	%	Exceptionally high (EHM) Number	%
Alone	4	4	100	3	50
	3	4	100	5	85
	2	4	100	6	100
	1	4	100	6	100
Peer	4	0	0	6	67
	3	3	75	9	100
	2	4	100	9	100
	1	4	100	9	100
Child	4	2	25	5	100
	3	5	63	5	100
	2	6	75	5	100
	1	8	100	5	100

Note.—Using a one-tailed chi-square approximation of the Kolmogorov-Smirnov two-sample test for unequal sample sizes, the following just significant levels result: AAM Alone vs. AAM Peer: $p \leq .02$; AAM Alone vs. AAM Child: $p \leq .05$; AAM Peer vs. AAM Child: $p \leq .40$, regardless of direction; EHM Peer vs. AAM Peer: $p \leq .08$; EHM Child vs. AAM Child: $p \leq .03$.

were available, the correlation between the two aptitude scores was .51 ($p \leq .0001$).

It will be remembered that a naïve S was used for a receiver in each trial in the PC. The peer receivers' test scores were classified into four levels on the same basis as the test scores of the transmitters. The scores of the peer receivers were then compared to the scores of the transmitters, comparing the scores of the EHM transmitters to the scores of their receivers and the scores of the AAM transmitters to the scores of their receivers. Again utilizing the Kolmogorov-Smirnov two-sample test, it was apparent that there was no difference between the test performances of the transmitters and receivers. In fact, only two out of 13 receivers failed to equal the performance level of their respective transmitters and these two were only one performance level lower. The results are the same if the comparison is made on the basis of the transmitters' verbal aptitude scores.

DISCUSSION AND CONCLUSION

The major issue that has been investigated is whether the expectation of transmitting a concept to another individual will affect its acquisition by the transmitter. The framework proposed earlier hypothesized that an individual might have a conceptual repertoire that is used for learning and problem solving (utilizing symbols from a symbolic repertoire) and a transmission vocabulary (that may be a subset of the symbolic repertoire) that is used for interpersonal communication. Since it was proposed that concepts difficult to describe might cause the greatest decrement in conceptual attainment, a task was designed that was difficult to verbalize but which allowed objective measures of performance.

The experimental design attempted to measure, in a control condition, the ability of Ss to acquire the concepts without the distraction of expecting to transmit. This condition was compared with two others: one where S might perceive the receiver, a scholastic peer, as having an equally developed conceptual repertoire and transmission vocabulary, and another condition where the transmitter might perceive the receiver, a 6-year-old boy, as having a less developed conceptual repertoire and transmission vocabulary.

A major problem resulted from the lack of variation in the dependent variable, the transmitters' test scores. To prevent suspicion of the transmitters in the Child Receiver condition, the task was designed so that the transmitters might infer that a 6-year-old could understand the concepts. However, to provide variation in the transmitter's performance, it is desirable to have a task difficult for S population. As most transmitters (20 out of 36) attained maximum scores on the test, it would appear that this task was easy for the population from which the transmitters were drawn.

After the transmitters were divided into two groups on the basis of their math aptitude scores (considered an independent measure of their ability at the task), the transmitters with above average math aptitude scores in the learning-with-expectation-to-transmit situations demonstrated significantly less learning than those in the learning-without-expectation-to-transmit situation.

In the Peer condition, the peer receivers had test scores equal to those of the transmitters. The Peer condition then corresponds to the $-L,T$ case described earlier. Assuming that the transmitters perceived that the receiver's conceptual repertoire and transmission vocabulary were equal to their own, it may be inferred that the very act of expecting to transmit led the transmitters to reduce their conceptual "tool kit" and utilize only those concepts they could transmit with ease and confidence in learning the subject matter. Given these assumptions, it may be concluded that the transmission vocabulary is significantly smaller than the symbolic repertoire.

Examining the results from the Child condition, there is only tentative evidence that there may be a greater potential for a reduction in learning than in the Peer condition. Considering the exploratory nature of this research, it would appear that this situation deserves further attention.

One important unanswered question is the effect on an S in a learning-with-expectation-to-transmit situation when the

receiver is perceived as having a more sophisticated and developed conceptual repertoire and transmission vocabulary. If there is a decrement in learning in this situation, then the evidence is very strong that a reduction occurs in the transmitter's conceptual tool kit and that it is mediated through the transmission vocabulary.

Another major issue is the possible interaction effect between the expectation to transmit and the expectation to be tested. It will be recalled that the Alone *S*s expected to be tested and the transmitters, who expected to transmit, did not expect to be tested. In essence, this issue concerns the effect on the transmitters when they expect to be evaluated on *both* their own test performance and the test performance of the receiver.

It should be mentioned that in the Finlay (1966) experiment the transmitters in the learning-with-expectation-to-transmit situation did *not* demonstrate a decrement in learning performance. Finlay's major result was the effect on the receiver's performance, which was significantly lower in the learning-with-expectation-to-transmit situation. Her result corresponds to the L, −T case, discussed earlier, while the present results appear to correspond to the −L, T case. It may be that this difference is attributable to the radically different types of material that the transmitters were asked to learn in the two experiments, a skill in the Finlay study and difficult to describe concepts in the present study. This raises another issue that deserves attention: What are the characteristics of subject matter that lead to these different effects in the learning-with-expectation-to-transmit situation?

REFERENCES

BRUNER, J. S., GOODNOW, J. J., & AUSTIN, G. A. *A study of thinking*. New York: Science Editions, 1962.

COHEN, A. R. Cognitive tuning as a factor affecting impression formation. *Journal of Personality,* 1961, **29,** 235–245.

DUNCAN, C. P. Recent research on human problem solving. *Psychological Bulletin,* 1959, **56,** 397–429.

FINLAY, F. Some effects of expectation to teach on pupil performance. Appendix I in Paul D. Reynolds, Certain effects of the expectation to transmit on concept attainment. Unpublished master's thesis, Stanford University, 1966.

GAGNÉ, R. M. Problem solving and thinking. *Annual Review of Psychology,* 1959, **10,** 147–172.

KENDLER, T. S. Concept formation. *Annual Review of Psychology,* 1961, **12,** 447–472.

LEEPER, R. Cognitive processes. In S. S. Stevens (Ed.), *Handbook of experimental psychology,* New York: Wiley, 1951. Pp. 730–757.

REYNOLDS, P. D. Certain effects of the expectation to transmit on concept attainment. Unpublished master of arts thesis, Stanford University, 1966.

SIEGEL, S. *Nonparametric statistics*. New York: McGraw-Hill, 1956. Pp. 127–136.

ZAJONC, R. B. The process of cognitive tuning in communication. *Journal of Abnormal and Social Psychology,* 1960, **61,** 159–167.

(Received February 1, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 3, 147–153

EFFECTS OF RULES OF THUMB ON TRANSFER OF TRAINING[1]

TRACY H. LOGAN[2] AND KENNETH H. WODTKE
Pennsylvania State University

A rule of thumb was added to an instructional program designed to facilitate transfer to problems by means of a more general principle. The presence of the rule of thumb produced a marked decrement in performance on the transfer tasks. Only 20% of the students who were given the rule achieved perfect scores on a transfer test, while 75% of the no-rules group achieved perfect scores. The transfer decrement occurred in spite of the fact that the students were given several didactic warnings indicating that the rule would not apply on the transfer problems. The poor performance of the rule-of-thumb groups may have resulted from their misuse or overgeneralization of the rule. The results were discussed in terms of the classical negative transfer paradigms and the effects of a persistent set on problem solving.

A number of previous experiments (Hendrickson & Schroeder, 1941; Hilgard, Irvine, & Whipple, 1953; Judd, 1908; Overing & Travers, 1966) have demonstrated that knowledge of a relevant principle facilitates transfer to problems which involve an application of that principle. Each of the above experiments have in common the fact that the principles taught were applicable to *all* of the transfer problems employed in the experiment. The principle had no exceptions. This situation is somewhat atypical of many instructional situations in which a number of principles of varying generality must be taught. The student's task is complicated by the need to learn exactly when each of several principles is applicable. Such a learning situation would seem to provide many opportunities for *negative* transfer.

A negative transfer situation may arise when rules of thumb[3] for solving problems are taught. In the present discussion a rule of thumb is simply regarded as a principle having only very limited generality. Such rules are often taught in conjunction with a more general principle. They are quite common in subjects such as mathematics, statistics, and science, and are justified as shortcut, time-saving procedures. The great difficulty with most rules of thumb is that they are usually only applicable to a very limited class of problems. Due to its

[1] The writers wish to express their appreciation to Bobby R. Brown who assisted in the data analysis, and to William Rabinowitz for commenting on an earlier draft of the paper. Partial support for the study was obtained under a United States Office of Education Project No. 5-85-074. The study was completed while the first author was on a National Science Foundation Faculty Fellowship.

[2] Now at Wabash College, Crawfordsville, Indiana.

[3] The terms "principle" and "rule" have not, in the writers' judgment, been carefully defined in the literature. The question of what is learned when a student learns a rule or principle has not been adequately answered, and progress in this area will be slow until an adequate answer to this question is obtained. The learning of a rule may consist of the chaining of a series of implicit or explicit verbal mediating responses which become linked to a certain class of problems. In problem solving, the process may be analogous to a series of if–then statements such as "If I am confronted with problem type X, then rule Y applies." The rule itself may be a relatively simple and straightforward statement of how to arrive at a solution to the problem as in most rules of thumb, or it may be a complex chain of mediating responses such as, "if A is true, then B must be equal to C, and therefore D equals . . . , etc." Rules may differ along a number of dimensions, thus, rules may vary in their generality or the variety of problems to which they apply, they may vary in the complexity of the mediational chains involved in applying the rule, or they may vary in their meaningfulness. Typically, the term "rule" is used when generality is limited and when the mediating response chains are relatively simple and straightforward. The term "principle" usually implies greater generality and greater complexity of if–then chaining. In the present study, the term "principle" is used to refer to a problem solving process which has wide generality, and which is fairly complex in the mediational processes involved. The term "rule of thumb" is used to refer to a rule which has less generality, and which involves much simpler if–then chaining.

limited generalizability, a rule of thumb can rarely mediate transfer from an original set of problems to a new set of problems. Much more generally applicable principles are needed to facilitate transfer to new problems. In spite of its limited generality, however, a rule of thumb may be invoked in the transfer situation because of the student's tendency to overgeneralize its use, and his failure to recognize its exceptions. Experience in the classroom often testifies to the fact that students appear to have a strong set to "blindly" apply a rule of thumb regardless of its applicability.

Schulz (1960) has suggested that the same principles which govern transfer in verbal learning may apply to some aspects of problem solving. If a general principle and a rule of thumb were both associated with a common class of problems, previous research on transfer would lead to the prediction that each would tend to generalize to similar problems on a transfer task. Any generalization of the limited rule of thumb would lead to incorrect solutions on any transfer problems to which it did not apply. The interference of the rule of thumb with the application of the more general principle would be expected to increase when the transfer problems were highly similar to the training problems, and when students were unable to determine the relevant differences among the problems. This latter expectation is consistent with the finding that inadequate stimulus predifferentiation leads to increased negative transfer (Ellis & Muller, 1964; Gagné & Baker, 1950). Finally, one might also predict from the basic research on transfer that the degree of interference from the rule of thumb would be a function of the degree of original learning of the general principle (Underwood, 1951). If the general principle was overlearned, one might expect less interference from the limited rule of thumb.

The present experiment sought to provide information on two questions related to the above expectations:

1. What is the effect on transfer of including a rule of thumb having only limited generality in an instructional program designed to teach a more general principle?

2. Will additional opportunities to practice using the general principle reduce the amount of negative transfer resulting from the knowledge of the rule of thumb?

METHOD

Subjects

The subjects (*Ss*) for the investigation consisted of 79 students from an introductory educational psychology class who volunteered for the experiment, and were given course credit for their participation. The experiment was conducted during the spring term, 1966. Although *Ss* represented a variety of different majors, a large proportion were enrolled in the teacher education program.

Materials

The topic of instruction chosen for the investigation was the concept of significant figures at the level of high school or introductory college physics. The concept of significant figures was especially suitable for a study of the effects of rules of thumb on transfer, since teachers and textbooks typically employ several rules and principles in teaching the concept.

Instruction was presented to *Ss* in the form of a self-instructional programmed text. There were three versions of the program corresponding to each of three treatment conditions. Version 1 consisted of 37 main-trunk frames, remedial frames, and several practice problems designed to teach *Ss* the general principle of significant figures. The *Ss* were taught the basic reasons for the loss of significance in figures. The *Ss* were shown that when a measurement is taken, the accuracy of that measurement is restricted by the limits of the scale's graduations. Thus, in a measurement of 9.53 centimeters taken from a typical meter stick, the figure 3 representing 3 hundredths of a centimeter is usually estimated visually since the meter stick is not graduated in hundredths of a centimeter. Since different investigators are likely to vary in their estimates of such figures, the results of calculations involving these estimates will also vary. For example, the following problem illustrates the errors that are likely to be made, and the procedure for arriving at the "significant" figures assuming that the error in the estimated value is $\pm.01$. The problem is to find the area of a rectangle whose sides are 9.53 and 8.67 centimeters long. One finds the largest and smallest values the length and width could have and from these the largest and smallest values the area could have, for example:

$$9.54 \times 8.68 = 82.8072$$

$$9.53 \times 8.67 = 82.6251$$

$$9.52 \times 8.66 = 82.4432$$

From these values it is seen that the only value

that is absolutely certain in the answer is 82. The tenths figure could range from about .4 to .8, but it could not be as little as .2 or as great as .9 so some degree of confidence can be placed in the tenths figure. Since the tenths place was only partially certain, the hundredths place would be completely uncertain and could take any value from 0 to 9. From this reasoning one arrives at 82.$\bar{6}$ as the correct number of significant figures with the tenths place marked to indicate only partial certainty in that figure. The above reasoning applies to all calculations involving measurements and therefore has wide generality.

All instruction was provided within the context of multiplication problems. Following the basic instruction on multiplication problems, Ss' ability to transfer this basic reasoning to addition and trigonometry problems was measured.

Version 2 of the program was exactiy like Version 1 except that immediately following the material on the basic principle of significant figures, Ss were given the rule-of-thumb section which included the following frame plus several illustrations of its application:

> A Quicker Way For Determining Which Figures In A Calculation Are Significant.
>
> The method you have used is excellent, but a bit slow. Here is a faster way. It is just a rule of thumb, and works only for products and quotients, not for addition.
>
> *Rule:* When multiplying or dividing, the result has just as many significant figures as the factor with the fewest significant figures.

This version is referred to as the "rule-early" program since the rule of thumb was given immediately following instruction on the basic principle, but prior to a short practice segment.

The third version of the program was exactly like the second version except that the rule of thumb was introduced at the very end of the program following the short practice segment. Version 3 is referred to as the "rule-late" version. The practice segment was designed to provide extra practice using the general principle prior to the introduction to the rule of thumb. Thus, the only difference between Versions 2 and 3 of the program was the placement of the rule of thumb in the instructional sequence; either before or after a short practice segment. Table 1 summarizes the instructional sequences in the three experimental treatment conditions, and the associations which the instructional segments had been designed to teach.

The two rule-of-thumb versions of the program also contained several explicit warnings concerning the exceptions to the rule of thumb. Upon introducing the rule of thumb, Ss were told: "It is just a rule of thumb and works only for products and quotients, not for addition." After introducing the rule, Ss were told:

> ...as with many rules, there are occasional exceptions where the rule gives an incorrect answer. Therefore you are strongly advised to check *any* rule result by using the basic reasoning of significant figures until you get a feeling for when the rule works and when it doesn't—say, at least for the next week or so.

Immediately preceding the transfer test Ss were again told: "Although you have not practiced these, you can reason them out. Just trust your brain." In addition to these cues, the rule of thumb itself contains a cue to the limits of its applicability, thus, the rule states, "When *multiplying* or *dividing*, the result has just as many significant figures as the factor with the fewest significant figures." (The italics were not included in the experimental program.)

A pretest and posttest were developed to measure Ss' ability to determine the correct number of significant figures in a calculation. The pretest contained five problems designed to assess Ss' prior knowledge of significant figures. The posttest contained five multiplication problems and eight transfer problems. The eight transfer problems consisted of five addition or subtraction problems, and three trigonometry problems. The transfer problems could not be solved correctly by the simple application of the rule of thumb, but they could be solved by applying the basic principle of significant figures which was taught in Stage 1 of the instructional program. The Kuder-Richardson formula 20 reliability of the total 13-item posttest was .78 for the total sample of 79 Ss. For the subsamples, the reliabilites for the five-item addition transfer test were .64, .75, and .72 for the no-rule, rule-early, and rule-late groups, respectively. The reliabilities at the three-item trigonometry transfer test were .53, .53, and .64 for the same three treatment groups, respectively. Some of the reliabilities are lower than those usually obtained for mathematics achievement tests. Although the lower reliabilities were in part due to the small number of items making up the subscores, the range of performance was also severely restricted since many Ss achieved perfect scores.

Procedures

A pool of 162 Ss in an introductory educational psychology class were given the significant figures pretest. Twenty-five Ss were uninterested in participating. Of the remaining Ss, only those whose pretest performance indicated minimal knowledge of significant figures (maximum scores of zero or one on addition and never more than two on multiplication with zero on addition) were asked to volunteer for the experiment. Seventy-nine students served as experimental Ss. The Ss were assigned at random to one of the three experimental conditions. Approximately 1 week after the administration of the pretest, Ss reported to a large classroom to complete the instructional program and the posttest. Two groups were tested on two successive evenings. Upon arriving at the experimental room, each S was given the version of the instructional program (either no rule of

TABLE 1

EXPERIMENTAL PARADIGMS AND ASSOCIATIONS ESTABLISHED DURING THE DIFFERENT EXPERIMENTAL STAGES

Stages of instruction	Condition 1—No rules	Condition 2—Rule early	Condition 3—Rule late
Stage 1—Pretest and basic instruction	Pretest Basic principle of significant figures in multiplication	Pretest Basic principle of significant figures in multiplication	Pretest Basic principle of significant figures in multiplication
Associations established	M→ Pr	M→ Pr	M→ Pr
Stage 2——Introduction to rule of thumb and practice segment	Practice M→ Pr	Rule given for multiplication M→ Ru Practice M→ Pr M→ Ru	Practice M→ Pr Rule given for multiplication M→ Ru
Stage 3—Measure transfer and multiplication posttest	Multiplication posttest Transfer posttest Addition Trigonometry	Multiplication posttest Transfer posttest Addition Trigonometry	Multiplication posttest Transfer posttest Addition Trigonometry

Note.—Abbreviated: M = multiplication problems; Pr = general principle; Ru = rule of thumb.

thumb, rule early, or rule late) depending on the experimental group to which he had been assigned. The seating arrangement in the room was staggered to prevent communication. Following completion of the instructional program each *S* was given the posttest containing the multiplication and transfer problems. After completing the posttest, the Rokeach Dogmatism Scale was also administered as part of a separate phase of the investigation. Following completion of this scale, *Ss* were thanked for their participation and dismissed. Approximately 2 weeks later, the nature and purpose of the experiment was explained to *Ss*.

RESULTS

Results were analyzed by means of one-way classification analysis of variance and Scheffé tests of individual comparisons. As a precaution against the possible effects of violations of assumptions underlying the analysis of variance, a Kruskal-Wallis one-way analysis of variance by ranks (Siegel, 1956) was also employed.

Means and standard deviations for transfer tests under each condition are shown in Table 2. Overall F tests were statistically significant for both the addition and trigonometry transfer scores ($F = 12$, $df = 2/76$, $p < .001$ and $F = 3.5$, $df = 2/76$, $p < .05$, respectively). The mean of 4.5 achieved by the no-rule group on the addition problems, and the mean of 1.6 achieved by the same group on the trigonometry problems indicates that the instructional program in the basic reasoning of significant figures was successful in producing considerable transfer to problems which were not specifically taught in the program. Eighteen of the 26 *Ss* in the no-rule group achieved *perfect* scores of five on the addition transfer test. A Scheffé comparison of the mean addition transfer scores of the two rule-of-thumb groups with the no-rule group indicates that the inclusion of the rule of thumb in the instructional program produced a considerable decrement in transfer. The mean addition transfer score of the no-rule group was significantly greater than the means of the rule-early and rule-late groups (p values were less than .05 and .01, respectively). Whereas 18 of 26 *Ss* in the no-rule group achieved perfect scores on the addition transfer test, only 20 of 53 *Ss* in the two rule-of-thumb groups achieved perfect scores.

The detrimental effect of the rule of thumb appeared somewhat weaker in the case of the trigonometry subtest. The Scheffé comparisons indicated that the no-

rule versus rule-late comparison was the only comparison which approached statistical significance at less than the .10 level, although the other comparisons were consistent in direction with the differences obtained for the addition test. The reduced statistical significance in the analysis of the trigonometry scores as compared to the addition scores may have been due to the smaller number of items and the resultant lower reliability of the trigonometry subtest.

The mean scores on the multiplication posttest (less one item which could not be solved by the rule-of-thumb method) were 2.9, 3.0, and 2.9 for the no-rule, rule-early, and rule-late groups, respectively. Thus, the differences in transfer of the three groups could not be ascribed to differences in achievement on the multiplication problems directly taught in the program.

The second question of interest in the study concerned the effects of the rule-practice sequence on transfer. It was hypothesized that an experimental condition in which a rule was given before a practice segment would have a more detrimental effect on transfer than a condition in which the rule of thumb was given after practice. It was expected that giving the rule of thumb late in the program would provide additional practice with the more general principle, thus reducing the amount of interference from the rule. This hypothesis was not confirmed. In fact, the only difference between the rule-early and rule-late groups which approached significance at less than the .10 level was in the direction opposite to that which was predicted. This unexpected finding may have resulted from the relatively short length of the practice segment (due to a limit on *S*s' time only three problems were employed in the practice segment) or from a recency effect of giving the rule of thumb just prior to the transfer test which may have increased its saliency during the test situation.

Since a large number of *S*s in the experiment achieved perfect scores on the transfer tests, the distributions of performance were generally negatively skewed. As a precaution against the possible effects of violations of the assumptions of normality of distributions, Kruskal-Wallis one-way analyses of variance by ranks were also computed (Siegel, 1956). The results of this analysis were consistent with the results of the parametric analysis of variance (for addition $\chi^2 = 283$, $p < .001$, $df = 2$; for trigonometry $\chi^2 = 292$, $p < .001$, $df = 2$).

TABLE 2

DESCRIPTIVE STATISTICS FOR THE THREE EXPERIMENTAL GROUPS ON THE TWO TRANSFER TESTS

Transfer test	Condition 1 no-rule		Condition 2 rule-early		Condition 3 rule-late	
	M	*SD*	*M*	*SD*	*M*	*SD*
Addition (5 problems)	4.5	.9	3.5	1.6	2.6	1.7
Trigonometry (3 problems)	1.6	.7	1.2	1.0	1.0	1.0

Note.—In Condition 1, $N = 26$; in Condition 2; $N = 27$; in Condition 3, $N = 27$

DISCUSSION

Although the present findings should be replicated and extended to other subject matters and other rules, the results suggest that teaching a rule of thumb of limited generality in an instructional program designed to facilitate transfer by means of a more general principle may produce considerable interference. Since many subject matters include a number of rules and principles varying in generality, teaching methods may have to provide opportunities to reduce the amount of interference in such situations.

Several tentative explanations may be offered to account for the results obtained. Schulz (1960) and Gagné (1964) have suggested that the same principles which account for simpler forms of associative learning may also occur in more complex forms such as principle learning or problem solving. Following this line of argument, it is interesting to ask whether the transfer paradigm employed in the present experiment in any way resembles the classical negative transfer paradigms employed in experiments on verbal learning? The paradigms of the present investigation are outlined in Table 1.

During Stage 1 of the experiment, *Ss* were taught how to determine the correct number of significant figures in a multiplication problem (M) by resorting to the general principle of significant figures (Pr). During Stage 2, *Ss* in the two rule-of-thumb groups were taught how to solve the same class of problems (M) by resorting to a much quicker rule of thumb (Ru) and were given practice either before or after the introduction of the rule of thumb. In the no-rule condition, *Ss* simply received the practice segment. (The practice segments were equated in the three conditions.) During the transfer test or Stage 3, all *Ss* were exposed to a new set of problems which could be solved by the general principle, but not be the rule of thumb. Although all *Ss* had instruction in Stage 1 which should have enabled them to solve the addition and trigonometry transfer problems, *Ss* in the two rule-of-thumb conditions had also acquired the competing response of the rule-of-thumb strategy. If one assumes that the didactic verbal warnings which were provided to warn *Ss* about the exceptions to the rule were ineffective (an assumption which seems well-founded in the present study), then *Ss* would be unable to determine the appropriateness of the rule-of-thumb solution during the transfer test. Under such conditions we would expect a high degree of interference in the form of competing responses from the rule-of-thumb solution which had been previously associated with the correct solutions of multiplication problems. Incorrect responses may result from *Ss*' tendency to overgeneralize the rule of thumb without regard for the distinctions between the multiplication, addition, and trigonometry problems. Since these three classes of problems are by outward appearances quite similar, the tendency of *Ss* to overgeneralize the use of the rule of thumb would not be surprising. The above interpretation suggests that in teaching rules of limited generality, it will be necessary to provide discrimination training so that the student will be able to make the relatively fine discriminations between the classes of problems to which the rule is applicable and the problems to which the rule is inapplicable.

It is interesting that the present paradigm, although not entirely parallel, bears some resemblance to a retroactive inhibition paradigm in which two competing responses are associated with the same or similar stimuli. One of the authors is conducting follow-up research to determine whether the detrimental effect of the rule of thumb is simply the result of overgeneralization of its use, or whether other factors are involved. A tentative analysis of *Ss*' errors on the transfer tests suggests that at least part of the effect is attributable to overgeneralizing the rule of thumb or rule misuse.

The effect of the rule of thumb appeared similar to the effect of a strong persistent set which seems to "blind" *Ss* to alternative solutions as reported in several classical experiments on set by Rees and Israel, (1935) and Luchins (1942). This apparently strong set to use the rule of thumb rather than alternative solution methods may result in part from its simplicity, ease of recall, and ease of application. It is also possible that the typical college student has been strongly preconditioned to use such rules through years of previous instruction which has emphasized rule-of-thumb solutions.

The present results also resemble the results of Wertheimer's (1959) work on teaching children to find the area of a parallelogram, and Katona's (1940) match-stick problems. Both of these investigators found that a condition in which *Ss* were taught a general principle facilitated transfer when compared to a condition in which a less general principle was taught or where *Ss* simply practiced problems. In Wertheimer's study, children who were taught to find the area of a parallelogram by a specific solution could not transfer to problems in which the parallelogram was placed in a different position, or to different geometric figures. Furthermore, when the children were presented with the new problems, they would often attempt to blindly apply the old inadequate method. However, there is one important differ-

ence in procedure between the Wertheimer and Katona experiments and the present experiment. In these earlier experiments, *Ss* were taught either by the method utilizing the general principle *or* by the specific solution method. It is not surprising that the methods produced differences in transfer since one group was never taught the relevant general principle. In the present experiment, *Ss* were taught *both* a method based on a general principle and a method based on the rule of thumb. The fact that the rule of thumb interfered with transfer even when an appropriate alternative solution was available to *Ss* emphasizes even more strongly the dangers of negative transfer which exist when such rules are available.

It might be argued that the verbal warnings which were given to *Ss* were simply not strong enough to call their attention to the limitations of the rule. This possibility must be recognized. The writers would agree that other warnings might have been included which might have had the desired effect on *Ss'* behavior. The warnings used in the present study were selected a priori by the writers, after considerable debate concerning their relative strength. Unfortunately, it was not possible in the present study to determine precisely the effect of the verbal warnings, since a rule-of-thumb group without the warnings was not included. However, it seems likely that didactic statements alone may not be very effective in establishing the discriminations needed to apply different principles to different problems.

Rules and principles may be a great aid in problem solving; however, the present study suggests that teaching rules with only limited generality may actually interfere with transfer. If these findings are replicated in other problem solving situations, one would either recommend that rules of limited generality not be taught where one has the alternative of teaching a more general principle, or that the instruction include opportunities for the student to learn when to use each of the alternative methods. It is possible that the additional instructional time required to teach the student to recognize the exceptions to a rule of thumb will offset the increased problem-solving efficiency in the limited class of problems to which the rule of thumb applies.

REFERENCES

Ellis, H. C., & Muller, D. G. Transfer in perceptual learning following stimulus predifferentiation training. *Journal of Experimental Psychology*, 1964, **68**, 388–395.

Gagné, R. M. Problem solving. In A. W. Melton (Ed.) *Categories of human learning*, New York: Academic Press, 1964. Pp. 293–317.

Gagné, R. M., & Baker, K. W. Stimulus predifferentiation as a factor in transfer of training. *Journal of Experimental Psychology*, 1950, **40**, 439–451.

Hendrickson, G., & Schroeder, W. H. Transfer of training in learning to hit a submerged target. *Journal of Educational Psychology*, 1941, **32**, 205–213.

Hilgard, E. R., Irvine, R. P., & Whipple, J. E. Rote memorization, understanding and transfer: An extension of Katona's card-trick experiments. *Journal of Experimental Psychology*, 1953, **46**, 288–292.

Judd, C. H. The relation of special training to general intelligence. *Educational Review*, 1908, **36**, 28–42.

Katona, G. *Organizing and Memorizing*. New York: Columbia University Press, 1940.

Luchins, A. S. Mechanization in problem solving: The effect of Einstellung. *Psychological Monographs* 1942, **54**(6, Whole No. 248).

Overing, R. L. R., & Travers, R. M. W. Effect upon transfer of variations in training conditions. *Journal of Educational Psychology*, 1966, **57**, 179–188.

Rees, H. J., & Israel, H. C. An investigation of the establishment and operation of mental sets. *Psychological Monographs*, 1935, **46**(6, Whole No. 210).

Schulz, R. W. Problem solving behavior and transfer. *Harvard Educational Review*, 1960, **30**, 61–77.

Siegel, S. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill, 1956.

Underwood, B. J. Associative transfer in verbal learning as a function of response similarity and degree of first-list learning. *Journal of Experimental Psychology*, 1951, **42**, 44–54.

Wertheimer, M. *Productive thinking*. New York: Harper's, 1959.

(Received March 30, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 3, 154-158

EFFECTS OF EFFORT ON RETENTION AND ENJOYMENT[1]

RUBY YARYAN BUENZ AND IRVING R. MERRILL
University of California, San Francisco

Dissonance theory predictions concerning the effects of effort on retention and enjoyment were tested in a human learning experiment. 60 Ss in 4 groups viewed identical films with interspersed sentence completion questions. 2 High-Effort Groups wrote 1-word answers plus a short essay for each question. 2 Low-Effort Groups wrote only 1-word answers. A High- and Low-Effort Group were tested after a 59-day retention interval; the other groups were tested after 40 days. All Ss completed tests of initial learning and retention plus some attitude questions. The results tend to support the hypothesis that greater effort leads to greater retention, and clearly confirm the hypothesis that greater effort results in greater enjoyment of the learning situation.

For years the variable of effort has had a confused status in learning theory (Lewis, 1965). The concept of effort has referred broadly to the magnitude of energy expended in responding to a learning situation. It emphasizes the sheer amount of energy expended in the entire situation surrounding the learning rather than on the specific repetition of a particular S-R sequence to be learned. In this difference involving the magnitude and direction of activity, effort can be distinguished from the other learning variables of practice and attention. Rather than facilitating a given behavior, as do practice and attention, effort is considered by most learning theories to be either a negative incentive or a variable that increases response inhibition. The evidence on the effects of effort during acquisition tends to support the view of effort having a negative or inhibitory value. Data regarding trials to extinction, however, raise serious questions for traditional views, since most learning theories fail to predict the effect of effort on resistance to extinction (Lawrence & Festinger, 1962).

As a solution to this theoretical impasse, the role of effort during learning and extinction has been explained in the context of dissonance theory (Festinger, 1961; Lawrence & Festinger, 1962). According to dissonance theory, the greater the effort that an organism is required to exert during the training situation, the more the organism resists extinction of the learned behavior. Effort is considered to be a deterrent to action, creating dissonance in the organism as it is induced to engage in effortful activity. Dissonance is reduced in the organism by developing "extra attractions" for the activity surrounding the learning or its consequences. Experiments using rats have demonstrated that increased effort leads to greater resistance to extinction regardless of the reward schedule employed during training. In these studies the development of "extra attractions" has been inferred to explain the phenomenon of greater resistance to extinction (Lawrence & Festinger, 1962).

The present experiment was conducted to test these dissonance theory predictions in a human learning situation. The specific experimental objectives were (*a*) to see if the dissonance theory predictions would hold for long-term retention in humans, and (*b*) to get some direct measure of positive affect associated with effort that would lend support to the postulation of the development of "extra attractions." In the application of dissonance theory three assumptions were made. First, the learning processes involved in trials to extinction in animals are analogous to measures of retention in humans, for both measures deal with how long in time a given learned

[1] Grateful acknowledgement is due Leon Festinger, Gretchen Franz, Jane Allyn Hardyck, Thomas S. Musser, and Mary Epling Smoke for their critical reading of the manuscript and valuable comments. For cooperation in the conduct of the experiment the authors thank Velena Boyd, Edith S. Deck, and Janice M. Payne.

behavior is maintained. This assumption seemed reasonable not only because of the temporal similarities of the two measures, but also because the present experiment was designed to simulate in a human setting the acquisition procedures used in extinction experiments with animals. Specifically, the material to be learned was presented in small segments, each segment requiring an effortful response. Inducing the participants to engage in repeated dissonant experiences during acquisition trials constituted the experimental manipulation. The second assumption was that human learning of abstract concepts involves principles that parallel (in a broad sense) the acquisition of simple perceptual-motor discriminations in animals. Third, it was assumed that the development of "extra attractions" is experienced by humans as enjoyment.

The hypotheses tested in the present experiment are as follows:

1. When people are required to exert a great deal of effort while they are learning, they will remember the learned material longer than people required to engage in little effort during the same learning situation.

2. Those people who exert a great deal of effort while they are learning will enjoy the learning conditions more than people who are required to exert little effort while learning the same material.

METHOD

A 2 × 2 design was used in the experiment. All subjects (*Ss*) were shown identical films. Interspersed throughout the films at periodic intervals were written questions requiring a one-word answer. Two groups of *Ss*, each tested initially at a different time, were required to write the one-word answer to each question plus a short essay justifying each answer; these *Ss* comprised the High-Effort Groups. Two Low-Effort Groups of *Ss*, each tested initially at a different time, viewed identical materials, and were required to write only the one-word answer to each question. After completion of the experiment all *Ss* were given a test of initial learning, and were asked several attitude questions. Several weeks later all the *Ss* were gathered together and given a retention test. The major dependent variables were (*a*) how much *Ss* remembered from the initial to the retention tests, and (*b*) how much *Ss* liked the experiment after they had completed the initial learning task. The details of the procedure follow.

Subjects

The *Ss* were 66 female students enrolled in a course in nursing. They were randomly assigned in equal numbers to the High- and Low-Effort Groups. Six *Ss* did not complete the experiment; the analysis was based upon 31 *Ss* in the High-Effort and 29 *Ss* in the Low-Effort Groups. Within each group, 17 *Ss* were retested after a 59-day test interval, and the remaining *Ss* were tested after a 40-day test interval.

Materials

The stimulus materials were two identical 56-minute films (television kinescope recordings) entitled *Recent Advances in Nursing Care of the Aged: Care of the Cerebral Vascular Accident Patient*. Each film included 20 sentence-completion-type questions, which appeared at about 3-minute intervals during the presentation.

Printed answer sheets were prepared to accompany the film. All sheets repeated each question in full and provided space for the appropriate written answer. For the Low-Effort Groups the questions were closely spaced; for the High-Effort Groups each page of the form contained only two questions, allowing space for a short essay after each question.

Most of the film questions involved factual answers that were difficult to dispute in an essay of justification. A typical question is exemplified by the following: *CVA is often caused by thrombosis, hemorrhage,* and ______. (answer: *embolus*) It should be noted that special attention was paid to the selection of questions for the essays. It was important to have questions in which effort could be manipulated without varying direct practice. Questions of a factual, indisputable nature were chosen for the essays of justification, because (*a*) the indicated answers were extremely difficut to enlarge upon and required considerable effort, and (*b*) the resulting essays, by necessity, contained information that was quite tangential to the critical material to be learned. For example, taking the question above, *S* was typically forced to base her justification on the fact that she was instructed in the film as to the answer and that is why she answered it as she did. Such a justification does not involve a repetition (practice) of the critical S–R connection to be learned. Evidence regarding the effectiveness of the manipulation will be presented below.

Procedure

Each group viewed identical films in separate but comparable rooms. Initial learning was tested immediately after the viewing sessions. The retest measuring the retention of all *Ss* was conducted at one time with the groups combined.

When *Ss* assembled for the initial learning

session they were informed that the purpose of the study was to evaluate their attitudes and reactions to the film, which was to be revised for broadcast television use. This deception was used to minimize suspicion that *S*s would be retested later.

Before the viewing, every *S* was given an answer sheet and a printed set of instructions. For the Low-Effort Group the instructions were to write the missing word on the answer sheet during the 15 seconds provided in the film after every question. Instructions to the High-Effort Group were to write the missing word and a short essay of justification for each answer; the presentation was interrupted after every question for about 2 minutes, so that the essay could be completed before the presentation resumed.

After the viewing, every *S* was handed a 40-item Initial Test to complete. They were told that test performances would be held in confidence by the experimenters and that course grades would not be affected. The High-Effort Groups took approximately 130 minutes and the Low-Effort Groups took approximately 95 minutes to complete this phase of the experiment.

At the retest session *S*s were told they were to complete a second evaluation of the film. They were given a 20-item Retention Test with the same instructions used for the previous 40-item test. When all *S*s completed the Retention Test, the true nature of the experiment was explained in detail.

Measurement

The 40-item Initial Test consisted of a random ordering of 20 learning questions that were related to the film questions, 3 attitude questions, 4 control questions, and 13 filler items. The filler items were not scored as part of the experiment. The control questions concerned situational factors that might affect learning. The attitude questions concerned enjoyment of the experiment, perception of effort expended, and perception of work performed. The learning questions covered the instructional content of the film. Each was a multiple-choice question with five alternative responses. The reliability (internal consistency) of the learning questions evaluated earlier on a population of hospital staff nurses, was $r = .66$. The 20-item Retention Test consisted of a different random ordering of the same learning questions.

The measure of retention for each *S* was the difference in number of correct responses between her Initial Test and Retention Test scores. The measure of enjoyment for each *S* consisted of her response to a direct question on a 5-point scale.

RESULTS

In the High-Effort Groups the total number of words written for the 20 essays by each *S* averaged 649. To evaluate the effectiveness of the manipulation, all *S*s were asked the following question on their perception of effort (with score values in parentheses): "How did you find this entire experiment now that you have completed it? Very effortful (5), Moderately effortful (4), Somewhat effortful (3), Slightly effortful (2), Not effortful at all (1)."

In the analysis of variance on perceived effort, only Effort Groups proved to be significant. The means for the High- and Low-Effort Groups were 3.00 and 2.24, respectively, indicating that *S*s in the High-Effort Groups perceived that they had exerted significantly more effort during the experiment than *S*s in the Low-Effort Group ($F = 5.40$, $df = 1/55$, $p < .05$).

The *S*s were also asked a question about their perception of work while they were viewing the films. When the two questions were combined into a "work-effort" index, parallel findings were obtained with a between-groups F ratio of 4.88 ($p < .05$). Correlations, however, between scores on the two questions were only $r = .38$ and .42 for the High- and Low-Effort Groups, respectively. Examination of the questions indicated that the wording of the question on work was ambiguous. Therefore, further analyses of perceived effort are reported using only the question on effort as an indicator.

On the sentence completion task (performed during the presentation of the films), the groups were very comparable with the mean number of correct responses for the Low-Effort Groups (16.14) slightly exceeding the mean for the High-Effort Groups (15.97).

The reliability coefficient for the internal consistency of the Initial Learning Test was $r = .48$, which was considerably lower than that for hospital staff nurses.

On the Initial Learning Test, as in the sentence completion task, the Low-Effort Groups scored slightly higher than the High-Effort Groups (Table 1). Analysis of variance on Initial Learning scores indicated that the Groups did not differ significantly in this respect ($F = .29$). The importance of these trends will be discussed in greater detail below.

Four control questions on (*a*) how well

Ss could see the films, (*b*) how well *Ss* could hear the films, (*c*) how helpful the questions were to *Ss* and, (*d*) how distracting the questions were to *Ss* indicated that the groups were very comparable. These questions were included to see if any spurious situational factors could account for any of the learning effects.

Differences in retention related to effort expended and test interval used were analyzed by a two-way analysis of variance for unequal frequencies in subclasses (Walker & Lev, 1953). The means for the Retention Test scores and the Retention change scores are presented in Table 1. The analysis of variance for Retention change scores showed no significant source of variance due to Effort Groups (F = 2.42), Test Intervals (F = .31), or the interaction between them (F = .38).

TABLE 1

Means for Initial Learning Test Scores, Retention Test Scores, and Retention Change Scores

Groups	N	Initial test scores[a]	Retention test scores[a]	Retention change scores[b]
High effort				
59-day interval	17	16.65	15.82	.82
40-day interval	14	15.43	15.21	.21
(Groups total)	(31)	(16.10)	(15.55)	(.55)
Low effort				
59-day interval	17	16.65	15.35	1.29
40-day interval	12	16.00	14.67	1.33
(Groups total)	(29)	(16.36)	(15.07)	(1.31)

[a] Based on total correct score for each subject.
[b] Based on difference between Initial and Retention Test scores for each subject. Smaller means indicate less change or greater retention.

To examine if effort was related to retention, correlation coefficients were computed between the retention change scores and the perception of effort ratings. Because the Test Interval and interaction effects were so small, the 59- and 40-day intervals were combined in each of the efforts groupings for this analysis. For the High-Effort Groups the correlation between perceived effort and retention was .28 ($p < .05$), whereas for the Low-Effort Groups perceived effort and retention were negatively related with a correlation of −.11 (*ns*). Following the hypothesis perceived effort was related to retention for the High-Effort Groups only.

Considering the analysis of variance and the correlations between retention and effort, the results provide positive but inconclusive evidence regarding the first hypothesis that people who exert a great deal of effort while learning remember the material better than people who engage in little effort during the same learning situation.

In order to test directly the second hypothesis, evaluating the effects of effort on enjoyment, all *Ss* were asked the following question on perceived enjoyment (with score values in parentheses): "Now that you have completed the experiment, what did you think of it? Enjoyed it very much (5), Enjoyed it moderately (4), Enjoyed it somewhat (3), Enjoyed it slightly (2), Did not enjoy it al all (1)."

Results of the analysis of variance for Perceived Enjoyment showed only Effort Groups to be significant (F = 4.29, df = 1/54, $p < .05$). The means for the High- and Low-Effort Groups were 2.19 and 1.70, respectively, indicating that the High-Effort Groups enjoyed the learning situation significantly more than the Low-Effort Groups.

To assess the relationship between effort and enjoyment, correlation coefficients were computed for the effort groupings. For the High-Effort Groups the relationship between the perception of effort and enjoyment was significant with an r of .27 (p = .05). For the Low-Effort Groups, however, enjoyment was not related to perceived effort (r = −.03, *ns*).

Following the prediction, perceived effort was related to enjoyment for the High-Effort Groups only. The results confirm the hypothesis that people who exert a great deal of effort while they are learning will enjoy the learning situation more than people who are required to exert little effort while learning the same material.

Discussion

The findings on enjoyment provide direct support for the hypothesis that dissonance, created by effort, is reduced in the organism by developing "extra attractions" for the learning conditions. This

experiment differs in two ways from an experiment by Aronson (1961), who studied the effect of effort on attractiveness. First, in Aronson's study, effort was confounded with reward, and the results were interpreted in terms of the interaction between dissonance and secondary reinforcement (a clear, unambiguous dissonance effect was not demonstrated). In this study reinforcement was held constant and enjoyment followed the dissonance-theory prediction. Second, Aronson's study concerned the expenditure of effort in obtaining colored containers, a simple, repetitive motor task. The present study was concerned with the expenditure of effort during the learning of unfamiliar material, demonstrating the effect of dissonance on enjoyment as it applies to an area of human learning.

The results on Initial Learning indicate that the manipulation was effective in producing effort that was irrelevant to the specific material to be learned. If the manipulation of effort had involved relevant practice, there should have been discernible improvement in initial-learning scores of the High-Effort Groups, but this did not occur. The initial-learning mean for the Low-Effort Groups was slightly greater than the mean for the High-Effort Groups. The same trend of the Low-Effort Groups scoring higher than the High-Effort Groups occurred with the mean scores on the sentence-completion task. If relevant practice had taken place, opposite trends would have been expected to occur.

The results on retention, although in the predicted direction, are somewhat equivocal. It appears that the comparatively low reliability of the test measuring the content learned from the films very likely increased the error variance and reduced the significance (see Walker & Lev 1953, p. 306). If improved test reliability would reduce the error term slightly, the first hypothesis that greater effort produces greater retention could be accepted.

The major implications of the experiment for educational practice would center on the activity or effortfulness of a student's response to educational materials (for example, materials including some of the elements of programmed instruction, especially those aspects dealing with the presentation of information in small segments followed by an effortful response). According to dissonance theory the more effort that the student is required to expend in the learning situation, the greater should be his enjoyment of the learning conditions and the greater should be his retention of the learned material.

In summary, the retention results can be considered suggestive, but cannot be considered a completely unequivocal test of the first hypothesis. The hypothesis concerning the effect of dissonance on enjoyment as it occurs in an effortful learning situation, however, was clearly confirmed in the experiment.

REFERENCES

ARONSON, E. The effect of effort on the attractiveness of rewarded and unrewarded stimuli. *Journal of Abnormal and Social Psychology*, 1961, **63**, 375–380.

FESTINGER, L. The psychological effects of insufficient rewards. *American Psychologist*, 1961, **16**, 1–11.

LAWRENCE, D. H., & FESTINGER, L. *Deterrents and reinforcement: The psychology of insufficient reward.* Stanford, Calif.: Stanford University Press, 1962.

LEWIS, M. Psychological effect of effort. *Psychological Bulletin*, 1965, **64**, 183–190.

WALKER, H. M., & LEV, J. *Statistical inference.* New York: Henry Holt, 1953.

(Received April 3, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 3, 159–163

TEST ANXIETY AND FEEDBACK IN PROGRAMMED INSTRUCTION[1]

PEGGIE L. CAMPEAU
Los Gatos, California[2]

The study evaluated effects upon criterion performance of S's test-anxiety level and the presence or absence of feedback in programmed instruction. 36 boys and 44 girls in the 5th grade scoring at high-anxious (HA) and low-anxious (LA) ends of the distribution on the Test Anxiety Scale for Children (TASC) were assigned to feedback (FB) and no-feedback (NO-FB) versions of a programmed lesson, yielding a 2 × 2 factorial design for each sex. An Anxiety × Feedback interaction was obtained for girls ($p < .025$), such that HA Ss did best under the FB condition ($p < .05$), and LA Ss surpassed HA Ss under the NO-FB condition ($p < .05$). No significant effects were found for boys. The results suggest that learning can be improved by adjusting programming procedures to adapt to individual differences in test-anxiety level, simultaneously reducing the negative effects of anxiety and capitalizing on the positive effects.

Investigators, including Mandler & Sarason (1952), Nicholson (1958), and a Yale team led by S. Sarason (Sarason, Davidson, Lighthall, Waite, & Ruebush, 1960) evaluated the effects of testlike conditions on the performance of high-anxious (HA) and low-anxious (LA) subjects (Ss) on conceptual learning tasks. The most consistent finding was that under the stress of testlike conditions, the performance of HA Ss was disrupted to a significantly greater degree than that of LA Ss. On the other hand, the direction of this difference was reversed when Ss were able to see whether or not they were improving from trial to trial (Mandler & Sarason, 1952), or when they received reassurance instructions in the nature of feedback (Sarason, 1957, 1958).

Reports of many investigators indicate negative correlations between measures of anxiety and scores on intelligence tests and college aptitude batteries (Feldhusen & Klausmeir, 1962; Hafner & Kaplan, 1959; McCandless & Castaneda, 1956; Phillips, King, & McGuire, 1959; Sarason, 1961a; Sarason & Mandler, 1952; Sarason et al., 1960). In a comprehensive review of findings related to the negative anxiety-ability relationship, Sarason et al. (1960) noted that even in studies where intelligence was controlled, there still were differences in learning rate between HA and LA groups.

Most of the anxiety research cited above employed an anxiety model which assumes that the elicitation and effects of anxiety depend on threat characteristics of the task situation and on S's anxiety level. The model stipulates that when anxiety has been learned as a response to situations involving intellectual achievement, anxiety feelings produced by comparable situations will elicit two types of responses: responses which are relevant to the task and lead to task completion, and responses which are task-irrelevant and interfere with task completion. One implication of the model is that in HA Ss, threat produces more task-irrelevant responses than task-relevant responses, and thereby disrupts performance. On the other hand, removing threat from the situation in which HA Ss must perform reduces task-irrelevant responses by reducing anxiety and thereby facilitates performance.

The suggestion for the present study was that different levels of anxiety may determine how insecure the learning situation can become without disrupting performance. Feedback was expected to be a vari-

[1] This study was conducted by the American Institutes for Research under Grant No. Title VII: 7-48-7670-229 for the United States Office of Education, Peggie L. Campeau, Principal Investigator. The experimental materials were developed by the investigator under a separate project (Contract No. OE 3-16-036) conducted by the Institutes for the United States Office of Education.

[2] Now a consultant to Dymedia, Inc.

able in programmed instruction style very likely to interact with anxiety level on the grounds that if feedback were omitted, testlike aspects of the program would be accentuated. Feedback (FB) and no-feedback (NO-FB) conditions were expected to have quite different effects on HA *Ss* as compared to LA *Ss*. Hypotheses and related theoretical considerations derived from the anxiety model and the research cited were as follows:

Hypothesis I. HA *Ss* would perform significantly better on a criterion test over the programmed lesson if they learned under the FB condition, as compared to the NO-FB condition. Feedback was expected to produce superior achievement for HA *Ss* based on two assumptions: (*a*) providing feedback would constitute a low-threat condition which would minimize task-irrelevant responses to anxiety by reducing that anxiety, and (*b*) task-relevant responses would be reinforced by providing feedback as confirmation.

Hypothesis II. HA *Ss* would perform significantly better than LA *Ss* on the criterion test, when both groups learned under the FB condition. Consistent with anxiety research already cited, providing answers to program frames in the FB condition was expected to reduce testlike aspects of the task so that criterion performance would be enhanced more for HA *Ss* than for LA *Ss*.

Hypothesis III. When both groups learned under the NO-FB condition, LA *Ss* would surpass HA *Ss* on the criterion test. Withholding feedback was expected to accentuate test-like aspects of the program. For HA *Ss*, anxiety produced by this condition would heighten the arousal of task-irrelevant responses, thereby disrupting their performance. On the other hand, LA *Ss* by definition are less likely to perceive testlike situations as threatening, so that their performance was not expected to be disrupted.

Hypothesis IV. Performance of LA *Ss* in the FB condition would not differ significantly from LA *Ss* in the NO-FB condition. The expectation of a no-difference finding for LA treatment groups was based on the assumption that, having very little anxiety to be manipulated, LA *Ss* would be less susceptible both to the facilitating effects of providing feedback and to the disrupting effects of withholding it.

METHOD

Subjects

The *Ss* were fifth graders at two elementary schools in the San Francisco bay area who scored at the high and low ends of the Test Anxiety Scale for Children (TASC).

Materials

A programmed instruction lesson on earth-sun relationships was prepared in two versions: one with feedback, one without. In all other respects, the two versions of the 193-frame program were identical. A pretest and posttest were prepared by random selection from a pool of 116 test items covering the content of the programmed lesson. The pretest contained 25 items; the posttest, 45 items.

Procedure and Design

The TASC, developed by Sarason et al. (1960), was administered to all fifth grade students at the two schools. Cut-off points for HA and LA groups corresponded approximately to the upper and lower 27% of the TASC distribution for each sex, yielding samples of 36 boys and 44 girls for experimental analysis. For HA girls, TASC scores ranged from 29 to 17; for LA girls, from 8 to 0. For HA boys, TASC scores ranged from 26 to 16; for LA boys, from 5 to 0. Samples were identified separately for girls and boys because previous research indicated sex to be a significant variable in experiments dealing with anxiety (Lunneborg, 1964; McCoy, 1963; Phillips, 1962; Sarason, 1961b, 1963; Sarason et al., 1960).

Nine days before the learning program was administered, scores on the pretest were obtained to provide measures of prior knowledge of the topic.

HA and LA *Ss* were assigned randomly to FB and NO-FB versions of the program in a 2 × 2 factorial design for each sex. To avoid giving the impression that certain students were being singled out for the experiment, even those individuals falling between the cut-off points in the TASC distribution were given one or the other version of the program to complete. The *Ss* progressed through the program at their own rates. Although mean time to complete the program was about 3 hours, 1½ school days were set aside for the experiment to allow *Ss* to take normal breaks for recess, lunch, and physical education, just as when working on class assignments.

The posttest was given to each *S* as soon as he completed his program. All *Ss* had time to complete the test.

The same test was readministered 19 days later

to obtain a measure of delayed retention. Again, all *S*s were allowed to complete the test.

To assure that instructions and procedures for all experimental measures and materials were equivalent at the two schools, (*a*) for every experimental session, *S*s assembled in a room large enough to accommodate the school's entire fifth-grade enrollment with adequate working space; (*b*) all experimental sessions were well monitored (approximately one monitor for every 12 *S*s); (*c*) the investigator always served as experimenter, assisted by the same monitors in all experimental sessions (teachers were never present), and (*d*) the same instructions were supplied to all *S*s either in written form or by the experimenter through announcements from a script.

RESULTS

Immediate Retention

A gain score (immediate retention test score minus pretest score) was calculated for each *S*. Because of the negative relationship between anxiety level and ability demonstrated in studies cited earlier, these gain scores were adjusted by covariance to control for differences in IQ as measured by the California Test of Mental Maturity (CTMM). Adjusted mean gain scores for *S*s in the different experimental treatments are given in Table 1. Scores for boys and for girls are shown separately.

For girls, analysis of covariance for adjusted gain scores yielded a significant Anxiety × Feedback interaction ($F = 5.61$, $df = 1/39$, $p < .025$). No significant effects were found for boys. Differences among adjusted means for girls' data were in the hypothesized directions. The significance of these differences was assessed by the critical difference method (Lindquist, 1953). For HA girls, the FB condition yielded significantly better performance than the NO-FB condition ($p < .05$). No other differences were large enough to achieve significance.

TABLE 1
ADJUSTED MEAN GAIN SCORES ON IMMEDIATE RETENTION TEST

Subjects	Program version: Feedback	Program version: No feedback
Girls		
High anxiety	25.85	16.47
Low anxiety	19.42	23.53
Boys		
High anxiety	17.14	18.12
Low anxiety	18.02	19.83

Note.—For girls, N = 11 in each cell; for boys, N = 9 in each cell.

TABLE 2
ADJUSTED MEAN GAIN SCORES ON DELAYED RETENTION TEST

Subjects	Program version: Feedback	Program version: No feedback
Girls		
High anxiety	23.09	14.85
Low anxiety	19.17	23.69
Boys		
High anxiety	21.72	20.32
Low anxiety	17.27	20.27

Note.—One or two *S*s were lost from some cells due to absences from the delayed retention test session. For girls, High-anxiety feedback, $N = 10$, High-anxiety no-feedback, $N = 11$, Low-anxiety feedback, $N = 9$, Low-anxiety no-feedback, $N = 9$; for boys, High-anxiety feedback, $N = 8$, High-anxiety no-feedback, $N = 9$, Low-anxiety feedback, $N = 7$, Low-anxiety no-feedback, $N = 9$.

Delayed Retention

Analyses similar to those made for the immediate retention test scores were performed on the data from the delayed retention test. The adjusted mean gain scores for delayed retention are presented in Table 2.

For girls, the Anxiety × Feedback interaction was significant ($F = 5.86$, $df = 1/35$, $p < .025$); again, no significant interaction was obtained for boys. HA girls under the FB condition surpassed the performance of HA girls under the NO-FB condition ($p < .05$), and LA girls surpassed the performance of HA girls under the NO-FB condition ($p < .05$).

DISCUSSION

Evaluation of the Experimental Hypotheses

For girls, three of the four hypotheses were supported: (*a*) HA *S*s did better under the FB condition than under the NO-FB condition; (*b*) LA *S*s did better than HA *S*s under the NO-FB condition, and

(*c*) LA *S*s did as well under the FB condition as they did under the NO-FB condition. The expectation that HA *S*s would do significantly better than LA *S*s under the FB condition was not supported. The results suggest that programming procedures be adapted to individual differences in test-anxiety level and sex. In this way, the negative effects of anxiety on learning could be reduced and the positive effects capitalized upon.

The contrasting findings for girls and boys deserve further comment. Evidence of sex differences in previous anxiety research cited earlier indicated that predictions about the interfering effects of anxiety were more clearly demonstrated for girls than for boys. In this study, the fact that girls' data supported experimental expectations while boys' data failed to confirm these predictions is generally consistent with earlier findings. Obviously, in generalizing the results of this study sex difference cannot be neglected.

Feedback during Learning and Facilitated Test Performance

Once the program had been completed and the test begun, all *S*s were in a "no-feedback" condition in that no answers were provided for checking their responses to test items. Since HA *S*s typically are at a disadvantage in a test situation, it might be expected that their performance would be disrupted during the criterion test. It is always difficult to differentiate between learning and performance. In this regard, further speculation may be in order as to why feedback during learning produced superior test performance for HA *S*s. It is suggested that the FB program allowed *S*s to *practice task-relevant responses.* That is, written responses were required to literally hundreds of questions throughout the program, and *S*s were reassured of the adequacy of their task-relevant responses by feedback. This procedure might be likened to taking a series of "practice quizzes" over program content as preparation for the criterion test. For HA *S*s most especially, feeling prepared to answer questions about program content could then have functioned to reduce disruptive effects of anxiety during the retention tests.

Suggestions for Future Research

In view of the fact that programmed instruction is finding increasing applications at all grade levels and in nearly all subject matters encountered in our schools, findings from this study suggest some worthwhile problems for future investigation. First, similar studies are needed in which feedback and no-feedback programs in different subject matters are administered to HA and LA learners at different grade levels. Data from such studies would permit greater generalization about the interaction between anxiety and feedback in programmed instruction, based on grade, subject-matter, and sex variables.

Second, more elaborate research is needed to take into account the teacher variable. When programmed instruction is integrated with conventional classroom teaching, the teacher will be another source of influence on student performance (Goldbeck, Shearer, Campeau, & Willis, 1962). In terms of the anxiety model outlined earlier, the nature of the teacher's response to student performance will help determine the degree to which classroom atmosphere is stressful or reassuring. The implication for future research is that differences in teacher responses to student performance should be studied in combination with programmed instruction variables (e.g., feedback) and with the learner's level of anxiety.

REFERENCES

FELDHUSEN, J. F., & KLAUSMEIER, H. J. Anxiety, intelligence, and achievement in children of low, average, and high intelligence. *Child Development,* 1962, **33**, 403–409.

GOLDBECK, R. A., SHEARER, J. W., CAMPEAU, P. L., & WILLIS, M. B. *Integrating programmed instruction with conventional classroom teaching.* San Mateo, California: American Institutes for Research, 1962.

HAFNER, A. J., & KAPLAN, A. M. Children's manifest anxiety and intelligence. *Child Development,* 1959, **30**, 269–271.

LINDQUIST, E. F. *Design and analysis of experiments in psychology and education.* Boston: Houghton Mifflin, 1953.

Lunneborg, P. W. Relations among social desirability, achievement, and anxiety measures in children. *Child Development*, 1964, **35**, 169–182.

McCandless, B. R., & Castaneda, A. Anxiety in children, school achievement, and intelligence. *Child Development*, 1956, **27**, 379–382.

McCoy, R. E. An experimental study of differences in test performance as a function of different intensities of chronic anxiety. *Dissertation Abstracts*, 1963, **24**, 2558.

Mandler, G., & Sarason, S. A study of anxiety and learning. *Journal of Abnormal and Social Psychology*, 1952, **47**, 166–173.

Nicholson, W. M. Influence of anxiety on learning: Interference or drive increment? *Journal of Personality*, 1958, **26**, 303–319.

Phillips, B. N. Sex, social class, and anxiety as sources of variation in school achievement. *Journal of Educational Psychology*, 1962, **53**, 316–322.

Phillips, B. N., King, F. J., & McGuire, C. Studies on anxiety: I. Anxiety and performance on psychometric tests varying in complexity. *Child Development*, 1959, **30**, 253–259.

Sarason, I. The effects of anxiety and two kinds of failure on serial learning. *Journal of Personality*, 1957, **25**, 383–392.

Sarason, I. Effects on verbal learning of anxiety, reassurance, and meaningfulness of material. *Journal of Experimental Psychology*, 1958, **56**, 472–477.

Sarason, I. Test anxiety and the intellectual performance of college students. *Journal of Educational Psychology*, 1961, **52**, 201–206. (a)

Sarason, I. Characteristics of three measures of anxiety. *Journal of Clinical Psychology*, 1961, **17**, 196–197. (b)

Sarason, I. Test anxiety and intellectual performance. *Journal of Abnormal and Social Psychology*, 1963, **66**, 73–75.

Sarason, S., Davidson, K., Lighthall, F., Waite, R., & Ruebush, B. *Anxiety in elementary school children: A report of research.* New York: Wiley, 1960.

Sarason, S., & Mandler, G. Some correlates of test anxiety. *Journal of Abnormal and Social Psychology*, 1952, **47**, 810–817.

(Received April 5, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 3, 164–168

TOKEN REINFORCEMENT OF ACADEMIC PERFORMANCE WITH INSTITUTIONALIZED DELINQUENT BOYS[1]

VERNON O. TYLER, JR. AND G. DUANE BROWN[2]
Western Washington State College

Court-committed boys ages 13–15 in a training school observed a daily television newscast. The following morning in school their teachers administered a 10-item true-false test based on program content; Ss were immediately shown their scores. After school, Ss were paid tokens redeemable for candy, gum, etc. During Phase I (17 days), Group 1 (9 Ss) received tokens contingent on test performance; Group 2 (6 Ss) received tokens on noncontingent ("straight salary") basis. During Phase II (12 days), Group 1 received noncontingent reinforcement and Group 2 contingent reinforcement. Hypothesis that test scores would be higher under contingent than noncontingent reinforcement was supported in both between- ($p < .05$) and within-S ($p < .005$) comparisons. Conclusion was that contingent token reinforcement strengthens academic performance.

Many educators prefer to motivate academic performance with "intrinsic" rather than "extrinsic" reinforcers; if used at all, they say, extrinsic reinforcers should be employed with caution (Marx, 1960). At the same time it is recognized that delinquent youngsters often have academic difficulty in the usual school situation (e.g., Bloch & Flynn, 1956; Briggs, Johnson, & Wirt, 1962). Since the IQs of delinquent youngsters may average well within the normal range (e.g., Tyler & Kelly, 1962), low motivation appears to be responsible for their poor school performance.

Various approaches have been suggested for motivating these "underachievers." As Birnbrauer, Wolf, Kidder, and Tague (1965) have indicated, these include (*a*) the use of "intrinscially reinforcing" materials which "are 'interesting,' 'meaningful'," etc., (*b*) "using materials and procedures which combine interest value and high probabilities of success," and finally, (*c*) "presenting social and/or symbolic reinforcers, e.g., teacher approval, grades, and stars." But as Birnbrauer et al., point out, none of these methods may be adequate for the retarded, school dropouts, and behavior problems. They suggest token reinforcement systems may be more effective. In such systems the tokens which are exchangeable for tangible reinforcers become generalized reinforcers (Skinner, 1953). A few examples of token reinforcement systems which have strengthened academic performance include studies with youngsters having reading difficulties (Staats, Staats, Schutz, & Wolf, 1962), retardates (Birnbrauer, Wolf, Kidder, & Tague, 1965), nursery school youngsters (Heid, 1964), and elementary school children (Michael[3]). However, work with delinquent youths appears to be quite rare. Cohen (Cohen, Flipczak, & Bis, 1965) has described a promising program for institutionalized delinquents. Of course, Slack (1960) and Schwitzgebel (e.g.,

[1] This study was conducted at Fort Worden Treatment Center (Washington State Department of Institutions, Division of Juvenile Rehabilitation), Port Townsend, Washington. Grateful appreciation is due Superintendent Gus Lindquist and Assistant Superintendent Robert H. Koschnick and Principal John Kanarr for their support and encouragement of this study; teachers Sam Rust, Jr., and William Harrison for their original thinking, which made this study possible, and for preparing and administering the tests; Cottage supervisor Allen Hodge and his staff Edith Smith and Lew Streit for administering the "token economy"; John D. Burchard and Don R. Shupe for their invaluable consultative services; Don Blood and B. L. Kintz for their comments on statistical procedures; and Patricia Soapes, Sara Burchard, Doreen Beazley, and Mary Wagner for collecting and compiling data.

A version of this paper was read at the Western Psychological Association, San Francisco, California, May, 1967.

[2] The authors are also at the Oregon State School for the Blind, Salem.

[3] J. Michael, personal communication, June 16, 1965.

Schwitzgebel & Kolb, 1964) have used operant techniques with delinquents, but not directly in the area of academic performance so far as is known.

For the present study it was assumed that many delinquent youngsters lack reinforced practice in the skills that result in teacher ratings of satisfactory performance. Apparently, the typical school situation does not provide the type of reinforcements necessary to strengthen these skills. The purpose of this study was to develop procedures for improving the academic functioning of a group of delinquent boys. This essentially involved setting up a "token economy" based on academic performance. More specifically, it was hypothesized that academic performance with contingent reinforcement will be superior to performance with noncontingent reinforcement in both between- and within-group comparisons.

METHOD

The subjects (*Ss*) in this study were 15 court-committed boys, 13–15 years of age who resided in a one-cottage living unit of a state training school. They attended school in their own self-contained classroom supervised by two team teachers. At 6 PM every evening, Monday through Friday, the television set in the cottage day room was turned on to the Huntley-Brinkley news broadcast. Youngsters were permitted, but not required, to watch the program; the only requirement was that all youngsters in the vicinity of the television set remain quiet so that those who wished to watch could do so. The following morning in school, *Ss* were administered a 10-item true-false test on the news program. The teachers wrote the questions the night before while watching the program. They wrote a new question every time there was a change of subject and two or three items to cover special subjects presented at the end of the program. The items were simple statements concerning the current events presented in the broadcast. Of course, this method meant the items were not standardized for difficulty. Immediately after administration, the tests were graded and the scores entered on a grade sheet which each student carried with him. Upon returning to the cottage, in the afternoon, those *Ss* on contingent reinforcement were paid in tokens according to the scores they had earned on the test; *Ss* on noncontingent reinforcement were paid a "straight salary." The tokens were redeemable for canteen items (candy, gum, etc.) and privileges in the cottage.

The *Ss* were paid the tokens according to a schedule designed by the experimenters (*Es*). The

TABLE 1

DESIGN FOR ADMINISTRATION OF CURRENT EVENTS TEST REINFORCEMENT

Subjects	Phase I (Days 1–17)	Phase II (Days 18–29)
Group 1	Contingent reinforcement	Noncontingent reinforcement
Group 2	Noncontingent reinforcement	Contingent reinforcement

Note.—For Group 1, $N = 9$; for Group 2, $N = 6$.

Es looked at each *S*'s scores on the true-false test for the 20 school days prior to the beginning of the experiment. Considering these data and *S*'s presumed level of motivation, a judgment was made as to what his schedule should be to maximize test performance; for example, if an *S* had been averaging 6 items correct, and had been earning approximately 20¢ a day in tokens, he would be given about 15¢ for 6 items correct, 20¢ for 7, 25¢ for 8, 27¢ for 9 and 30¢ for 10 correct. The goal was to let each *S* earn his previous average "income" with a slight improvement in performance and even more with greater improvements. The *Es*' judgments were influenced by a subjective assessment of *S*'s level of aspiration, tolerance for frustration, and limitations on the research budget for reinforcers.

Twenty *Ss* were randomly assigned to groups in the design outlined in Table 1. However, because of the rapid turnover of population in a crowded institution, some *Ss* left the institution prior to completion of the study resulting in unequal *Ns* in the two groups.

In Phase I, *Ss* in Group 1 were placed on contingent reinforcement and *Ss* in Group 2 on noncontingent reinforcement (paid 21¢ a day regardless of how well they did on the test). In Phase II, Group 1 was placed on noncontingent reinforcement and Group 2 on contingent reinforcement. A counterbalanced design was necessary to compensate for uncontrolled variability in the difficulty of the tests from day to day.

Although Group 1 (mean age 15.6) averaged a year older than Group 2 (mean age 14.6), both groups were functioning in the low average IQ range (mean IQs 94 and 98, respectively).

In addition, problems in data collection should be mentioned. On some occasions *Ss* were absent from school and could not take the tests. Because the tests were not equated for difficulty from day to day, the problem of missing data was a serious one. Only data for the days on which scores from at least 12 of the 15 boys were available were included for analysis. Missing scores for each *S* were replaced with the *S*'s mean score for the phase. From Phase I, data are reported from 17 of the 27 days on which tests were administered; from Phase II, data are reported from 12 out of 29 days. As is apparent, it was necessary to eliminate large quantities of data in order to make comparisons in

which most of the Ss of both groups were represented.

RESULTS

Mean daily test scores for both groups for Phases I and II over the 29 days reported are presented in Figure 1. Means for each phase are also included. The data show a clear pattern: during Phase I, Group 1 surpassed Group 2 on 15 out of 17 days; during Phase II, Group 2 surpassed Group 1 on 9 out of 12 days. Reversals when they did occur were quite small in contrast to the predicted differences between groups. The irregular, spiked form of the curves suggests that the tests varied a good deal in difficulty level from day to day as was expected. The nearly parallel form of the two curves indicates the groups responded to these variations in difficulty in a highly consistent, reliable fashion.

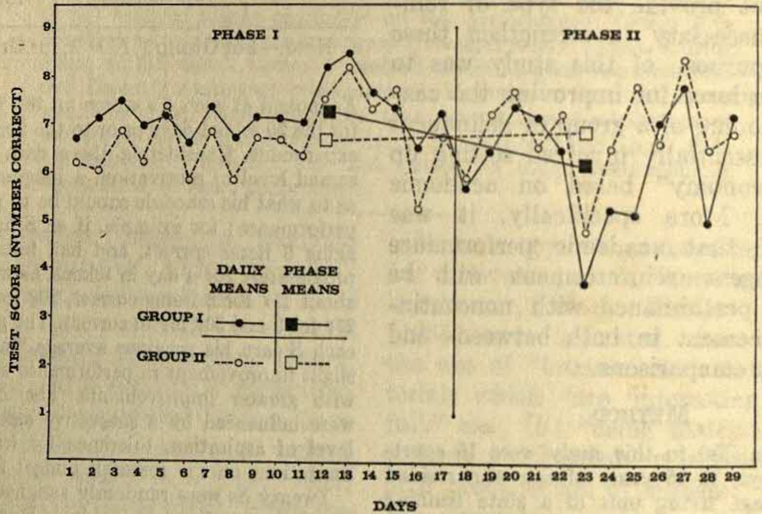


FIG. 1. The effect of contingent and noncontingent token reinforcement on true-false test performance.

TABLE 2

ANALYSIS OF VARIANCE OF SUBJECT MEAN[a] TEST SCORES UNDER CONTINGENT AND NONCONTINGENT REINFORCEMENT

Source	df	MS	F
Between Ss			
B (Groups)	1	.02	
Error (b)	13	.43	
Within Ss			
A (Phases)	1	2.60	7.43*
AB	1	2.49	7.11**
Error (w)	13	.35	

[a] Analysis based on two values from each subject: the mean of his 17 daily scores from Phase I and the mean of his 12 scores from Phase II.
* $p < .025$; two-tailed.
** $p < .0125$; one-tailed.

The S means for each phase were treated with a Lindquist (1953) Type I design analysis of variance. The summary of this analysis in Table 2 indicates there was no difference between Groups 1 and 2 (B comparison), but that the difference between Phases I and II (A comparison) and the interaction between Groups × Phases (AB comparison) were both significant ($p < .025$ and $p < .0125$, respectively). No difference was expected in the B comparison because of the counterbalancing of treatments. The difference between phases may be attributed to uncontrolled day-to-day variability in the difficulty level of the tests. The interaction effect indicates that under contingent reinforcement, performance was at a significantly higher level than under noncontingent reinforcement. Since the direction of the interaction effect was predicted, the probability value was halved (one-tailed test).

While the within-S variances against

which the interaction effect was tested were not significantly heterogeneous, some question could be raised about the normality of distributions of within-S difference scores. To avoid the assumptions of homogeneity of variance and normality of distributions and to study individual S performance, the data were subjected to nonparametric treatment. Mean scores for each S for each phase (same data as for analysis of variance) were classified as to whether the trend in the data supports (+) or does not support (−) the prediction that each S will perform at a higher level when token reinforcement is contingent on his test score than when it is not. Twelve of the 15 Ss did better under contingent reinforcement; only two Ss did worse. The Wilcoxon matched-pairs signed-ranks test (Siegel, 1956) was applied to these data yielding a highly significant T of 11 ($p < .005$; one-tailed test).

The between-groups effects were also tested for each phase separately, using the Mann-Whitney U test for independent measures (Siegel, 1956). As predicted, using one-tailed tests, during Phase I Group 1 surpassed Group 2 ($U = 13$, approaches significance at the .05 level); during Phase II, Group 2 surpassed Group 1 ($U = 12, p < .05$).

Discussion

Both between-groups and within-groups data clearly indicate that contingent reinforcement was associated with higher test performance than when reinforcement was noncontingent. This pattern emerged in spite of the use of quickly prepared unstandardized test items which varied considerably in difficulty from day to day and in spite of unstable conditions such as the shifting institutional population.

Moreover, this pattern appeared and was maintained with consistency over a 12-week interval. This would suggest more than a transitory effect, more than delinquents "playing games" with the program or a novelty that wore off.

While the effect of the contingent reinforcement is statistically significant, the practical educational significance appears limited at this point. The Ss on contingent reinforcement averaged less than one test item better performance than when they were on noncontingent reinforcement. On the other hand, it should be noted that Ss were attending small classes (10 students per teacher) led by teachers who in Es' judgment were about the most competent they had ever seen. These teachers had a knack with obstreperous youngsters; they knew how to discipline them and yet they were quite skilled and ingenious at devising methods of exciting the interests of even the most apathetic youngster. Thus the token reinforcement was tried against the severe competition of undoubtedly powerful social reinforcements supplied by these teachers.

That the reinforcement showed an effect in addition to what was generally regarded as an effective instructional program is further evidence of the importance of tangible reinforcers with delinquent and disadvantaged youngsters. It is doubtful that the tokens would have been this effective in a prosperous urban junior high school in which the youngsters were satiated with tangibles, enjoyed school, and were achieving "success" in the middle-class culture.

Replication of this study with more precise controls would more clearly demonstrate the effectiveness of this procedure. These controls should include an unreinforced control group and test items constructed to be more nearly equal in difficulty. Previous efforts by the investigators to produce improved academic performance with token reinforcement showed no results, presumably because of inadequate controls, particularly with regard to the measurement of the criterion.

Ultimately, of course, efforts must be made to "wean" youngsters from token reinforcers and link academic performance to the more traditional reinforcers such as social approval and perhaps even the "intrinsic" reinforcement of work "for the joy of the working [Kipling, 1896]." However, the results of the present study are encouraging and suggest that many young-

sters who are uninterested and antagonistic toward school work can learn that school work can "pay off."

REFERENCES

BIRNBRAUER, J. S., WOLF, M. M., KIDDER, J. D., & TAGUE, C. E. Classroom behavior of retarded pupils with token reinforcement. *Journal of Experimental Child Psychology*. 1965, **2**, 219–235.

BLOCH, H. A., & FLYNN, F. T. *Delinquency, the juvenile offender in America today*. New York: Random House, 1956.

BRIGGS, P. F., JOHNSON, R., & WIRT, R. D. Achievement among delinquency-prone adolescents. *Journal of Clinical Psychology*, 1962 **18**, 305–309.

COHEN, H. L., Flipczak, J. A., & BIS, J. S. *CASE Project: Contingencies Applicable for Special Education, Brief Progress Report*. Silver Spring, Md.: Institute of Behavioral Research, 1965.

HEID, W. H. Nonverbal conceptual behavior of young children with programmed material. Unpublished doctoral dissertation. University of Washington, 1964.

KIPLING, R. *The seven seas*. New York: Appleton, 1896.

LINDQUIST, E. F. *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin, 1953.

MARX, M. H. Motivation. In Marie R. Liba (Ed.), *Encyclopedia of educational research*. New York: Macmillan, 1960. Pp. 888–901.

SCHWITZGEBEL, R., & KOLB, D. A. Inducing behaviour change in adolescent delinquents. *Behaviour Research and Therapy*, 1964, **1**, 297–304.

SIEGEL, S. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill, 1956.

SKINNER, B. F. *Science and human behavior*. New York: Macmillan, 1953.

SLACK, C. W. Experimenter-subject psychotherapy: A new method of introducing intensive office treatment for unreachable cases. *Mental Hygiene*, 1960, **44**, 238–256.

STAATS, A. W., STAATS, C. K., SCHULTZ, R. E., & WOLF, M. M. The conditioning of textual responses using "extrinsic" reinforcers. *Journal of Experimental Analysis of Behavior*, 1962, **5**, 33–40.

TYLER, V. O., JR., & KELLY, R. F. Cattell's HSPQ as a predictor of the behavior of institutionalized delinquents. *Psychology Research Report No. 2.* Port Townsend, Wash.: Fort Worden Diagnostic & Treatment Center, 1962.

WINGO, G. M. Methods of teaching. In Marie R. Liba (Ed.), *Encyclopedia of educational research*. New York: Macmillan, 1960. Pp. 848–861.

(Received April 12, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 3, 169-175

RELEVANT AND IRRELEVANT VERBALIZATION IN DISCRIMINATION AND REVERSAL LEARNING BY NORMAL AND RETARDED CHILDREN[1]

NORMAN A. MILGRAM AND JAMES S. NOCE
Catholic University of America

Children age 7 and mental retardates of comparable MA were administered a discrimination learning and reversal shift task under varying conditions of relevant and irrelevant verbalization. When compared with a condition of no-verbalization, verbalization of the relevant class of cues facilitated performance, verbalization of 1 relevant and 1 or more irrelevant dimensions was also facilitative, but to a lesser degree, and verbalization of irrelevant dimensions only interfered. These effects were more evident in discrimination learning than in reversal shift and more evident in normals than in retardates. Results were interpreted as supporting the notion that verbal labels may direct attention toward or away from criterial dimensions.

The research literature on the effect of verbalization on discrimination learning and reversal shift in children has not yielded definitive results. Spiker (1963) reported that learning verbal labels for stimuli led to more rapid acquisition of relevant discriminations. Spiker did not utilize a reversal shift paradigm, but on the basis of his explanatory concept of enhanced cue distinctiveness one would predict that verbal labels referring to criterial cues should facilitate reversal shift as well as the original discrimination learning.

O'Connor and Hermelin (1959) confirmed a diametrically opposite prediction for a group of trainable mental retardates, namely that verbalization of relevant cues actually interfered with reversal shift. They argued that verbal connections in subjects (*S*s) below MA 6 whether normal or retarded would operate in a rigid stereotyped manner so that verbalization of a criterial cue, (e.g., "the big one") which was correct during discrimination learning would interfere with acquiring the new or reversed reponse (the smaller stimulus) during reversal shift. They acknowledged, however, that *S*s of higher MA, whether normal or retarded, could utilize verbal connections with greater flexibility and might well profit from overt verbalization of relevant cues in solving reversal shift problems.

Kendler, Kendler, and Wells (1960) reported that verbalization by 4-year-old children of relevant or irrelevant cues during a block of trials between original discrimination learning and reversal shift had no effect on the latter. In a second study (Kendler & Kendler, 1961), children age 4 and 7 were trained to give relevant or irrelevant responses during original discrimination learning. Results indicated that irrelevant verbalization significantly hindered reversal shift at both age groups, while relevant verbalization facilitated reversal shift for the 4-year-old children only. These investigators attributed the failure of the 7-year-olds to benefit from overt verbalization of relevant cues to spontaneous covert verbalization of relevant cues, consequently overt verbalization of relevant cues would not make any difference.

It appears reasonable to assume, however, that if the interfering effect of irrelevant verbalizations can be demonstrated, the opposite effect should also obtain. Failure to demonstrate verbal facilitation may have been due to a methodological defect in the Kendler paradigm. The reversal shift

[1] A portion of this paper was submitted by the junior author as a master's thesis to the Department of Psychology and Psychiatry, Catholic University of America, 1966. The senior author served as major professor on the thesis committee and was Principal Investigator of National Institute of Mental Health Grant MH 08488 which supported the overall study.

required was relatively easy for the 7-year-olds and was attained almost immediately. Children merely had to shift from one cue of a brightness or size dimension to the other cue. If the stimuli were varied on more than two dimensions, discrimination learning and reversal shift would be more difficult and the facilitation of overt relevant verbalization would then be demonstrated. To increase task difficulty four pairs of stimuli were utilized in the present study, each pair varying simultaneously along two cue values of each of three dimensions (size, brightness, shape) for example, a small black square versus a big white circle.

The Kendler paradigm, moreover, appears to have confounded reversal and nonreversal shift. Children were instructed to verbalize a cue within a dimension that was to become relevant or irrelevant only in the subsequent shift condition. Children in the relevant condition were trained to verbalize brightness and retained the brightness dimension with reversed cues on the second task. The other children were trained to verbalize size, which was also a correct guide to mastering the original discrimination task, but which became irrelevant during reversal shift when only brightness was consistently reinforced. From the point of view of these *S*s, the second task consisted of a nonreversal shift, that is, shifting from size to the brightness dimension. Viewed in this manner, the Kendler results may be interpreted merely as showing that children age 7 find reversal shift easier than nonreversal shift rather than as evidence for the interfering effects of irrelevant verbalization.

One way to meet this methodological objection is to require verbalization whose irrelevance to the criterial dimension is evident to the child each time he utters the verbalization. Let us assume that the children are learning to discriminate size and that one group has been instructed to designate response choices by verbal labels referring to the dimension of shape. This dimension is irrelevant throughout since a circle stimulus and a square stimulus are presented on each trial and shape is reinforced on a choice basis. One would predict that children employing verbal labels of the shape dimension are less likely to attend to the criterial size dimension than a base-line condition of no verbalization and would require more trials to reach criterion either in the original discrimination phase or the subsequent reversal shift phase. By contrast, children instructed to employ size designations are more likely to attain and to reverse the size discrimination than *S*s in the base-line condition.

It is an open question whether employing relevant and irrelevant verbal designations of response choices cancels itself out and is not significantly different from the no-verbalization condition or whether it exercises a significant effect in either direction. The present investigators favored the notion that mixed verbalization, relevant and irrelevant designations, would facilitate performance by directing attention to the criterial dimension being reinforced consistently. The directing of attention to irrelevant dimensions would be of less magnitude since these dimensions are reinforced only on a chance basis. The overall effect would thus be facilitative.

The present study attempted to answer these questions in an experimental design comparing two groups of children of different IQ level, but of comparable MA. In reviewing the experimental literature on the relative ease of discrimination learning by normals versus educable retardates of comparable MA, Stevenson (1963) states that no definitive conclusion has been reached. The relative ease of reversal shift also remains an unanswered question. In accord with the Lewin-Kounin theory of rigidity or the view of Luria (1957) and Reese (1962) on verbal mediation deficiency, one would predict an inferior performance by retardates, but Plenderleith (1956) and Stevenson and Zigler (1957) report equivalent performance by retardates and normals of comparable MA.

METHOD

Subjects

Normal *S*s were 96 children with a mean CA of 7.5 years and a range from 7 years up to but not

TABLE 1
Mean CA and MA for Retardates

Condition	CA		MA	
	M	*SD*	*M*	*SD*
1DR	15.7	4.5	6.7	1.5
2DR	15.6	4.0	7.2	1.2
3DR	14.4	2.3	6.8	0.8
C	16.7	4.0	7.0	1.0
1D-IR	13.7	2.8	6.8	0.8

including the eighth birthday. These children were selected from day camps located in the suburban Washington, D. C., area and were randomly assigned, 16 each to one of six experimental conditions. The retardates were 70 residents of the Children's Center of the District Training School, Laurel, Maryland.[2] The scarcity of suitable retardates restricted their participation to five major conditions, and they were randomly assigned to one of each. Mean CA and MA scores based on a recent Peabody Picture Vocabulary Test of the 14 *Ss* assigned to each condition are presented in Table 1. The mean CA ranged from 13.7 to 16.7 and the mean MA score ranged from 6.7 to 7.2. Overall, there were no significant differences in CA or MA between retarded *Ss* assigned to the various experimental conditions.

Stimulus Materials

Stimuli were two-dimensional squares and circles mounted in pairs on cardboard rectangles. The stimuli varied in size (1 inch and 3 inches) and brightness (black and white). The eight resulting stimuli were presented in four combinations of pairs such that each member of the pair differed from the other in all three dimensions (size, brightness, and shape).

Procedure

Pretraining in verbal labels. All children were tested individually by the second investigator. He explained that *S* was going to play a game but must first become acquainted with the game materials. The *S* was shown the stimuli in pairs and was instructed to designate all stimuli by one, two, or three dimension labels. Depending upon conditions, *S* employed designations of size ("the big one," "the little one"); brightness ("the black one," "the white one"); shape ("the circle," "the square") or a combination of these. In the experimental conditions which combined relevant and irrelevant dimensions or utilized more than one irrelevant dimension, *Ss* were trained to use such combined designations as "the big black one," "the white circle," "the little square," "the big black square," etc. Each *S* was required to reach a criterion of once through all eight stimuli giving each its proper designation. In the discrimination and reversal shift phases each *S* always indicated response choice between each pair of stimuli by employing only the designation peculiar to his pretraining experience. The control group who were not to employ verbal designations at all but to indicate discrimination choice by pointing were exposed to the eight stimuli the average length of time required by the various verbal groups to complete the pretraining phase (2 minutes). These *Ss* were commanded to look carefully at each card, but were neither given verbal labels nor permitted to employ their own verbal designations.

Discrimination learning. Each *S* was informed that the object of the game was to acquire a plastic poker chip each time he chose the correct of two presented stimuli. He was further told that after the game he would be able to trade in his pile of poker chips for a candy bar. The four pairs of stimuli were presented in a systematic alternation such that no one pair of stimuli appeared more than twice in succession and across all trials the correct cue appeared as often on the right as on the left. Following each trial the stimulus placards were raised. If a poker chip was found under the placard that was chosen, it was then given to *S* both as reward and feedback that he had made the correct choice. Between trials a vertical screen was imposed while the wells were baited and then covered by the next pair of stimuli. Criterion of successful learning was 9 out of 10 correct choices to a maximum of 72 trials.

Reversal shift. These trials followed the learning trials without any further instruction or ostensible variation in procedure. The difference was that chip reinforcements were now given for the reverse of the previously reinforced dimension. The *Ss* who failed to reach criterion on the original discrimination were not exposed to the reversal condition for the reason that one cannot be expected to reverse a discrimination that one has not acquired. Reversal shift continued until criterion of 9 out of 10 correct trials or a maximum of 72 trials was reached.

Experimental Schema

The relevant dimensions were size and brightness, while shape was never relevant. One dimension only was relevant for a given *S* in a given condition, yet all three dimensions were available simultaneously in the presentation of any pair of stimuli. The experimental conditions were as follows: (*a*) Verbalization of the one relevant dimension, size or brightness (1DR). (*b*) Verbalization of one relevant and one irrelevant dimension, the coupling of size or brightness with the other or with shape (2DR). (*c*) Verbalization of one relevant and two irrelevant dimensions, size or bright-

[2] The authors wish to express their deepest appreciation to the staff and residents of the Children's Center, District Training School, Laurel, Maryland, for their cooperation.

ness with the other and shape (3DR). (*d*) Verbalization of one irrelevant dimension, either size, brightness or shape (1D-IR). (*e*) Verbalization of two irrelevant dimensions, either size or brightness and shape (2D-IR). (*f*) A control or nonverbalization condition (C).

RESULTS

Two types of data were analyzed, frequencies of *S*s successfully reaching criterion and mean number of trials. Since the variances of trial scores varied widely between conditions these scores were routinely subjected to a square root transformation prior to analysis. All statistical tests reported were two-tailed. Frequency and transformed trial scores in the discrimination learning phase are presented in Table 2. It is noted that the distribution of frequencies by condition followed prediction and was nearly identical for normal and retarded groups. When group frequency data were collapsed, 1DR was significantly superior to C ($\chi^2 = 7.07$, $p < .01$) and 1D-IR was significantly inferior to C ($\chi^2 = 4.43$, $p < .05$). The conditions combining relevant and irrelevant dimensions, 2DR and 3DR, were equivalent to C and were also significantly superior to 1D-IR (and 2D-IR for normals also).

Analysis of variance of transformed trials for all *S*s in five experimental conditions (data on 2D-IR for normals were not included) yielded a highly significant *F* ratio ($F = 10.99$, $df = 4/130$, $p < .01$) for the condition effect. The array of means constituted a near perfect hierarchy in

TABLE 2
MEAN TRANSFORMED TOTAL TRIALS OF DISCRIMINATION LEARNING IN NORMALS AND RETARDATES

Condition	Normals: Number of subjects reaching criterion (Maximum 16)	Normals: *M*	Normals: *SD*	Retardates: Number of subjects reaching criterion (Maximum 14)	Retardates: *M*	Retardates: *SD*
1DR	16	3.79	0.61	14	3.73	0.72
2DR	13	5.07	2.00	13	4.63	1.54
3DR	12	5.64	2.14	10	5.76	1.95
C	11	6.39	1.91	11	5.44	2.05
1D-IR	8	6.40	2.23	6	6.95	2.11
2D-IR	6	7.95	0.97			

the predicted direction and *t* tests of differences between conditions were significant for 1DR > 2DR ($p < .05$), 2DR > C ($p < .05$), and in turn, C > 1D-IR ($p < .05$). The effect of IQ level and the interaction were not significant.

In the above analysis differences in mean trials between conditions were confounded with differential frequencies of *S*s successfully reaching criterion, since nearly all *S*s reached criterion in some conditions and only half in others. We may ask whether the predicted differences between conditions also obtain when we consider the trials of *only those S*s *who successfully* reached criterion. Mean transformed scores were computed in this manner and the resulting array was nearly identical in order to the earlier array for total trials. For normal children, 1DR = 3.79, 2DR = 4.28, 3DR = 4.69, C = 5.44, 1D-IR = 4.31, 2D-IR = 7.06. For retardates the corresponding means were 3.73, 4.33, 4.67, 4.61, and 4.90. An overall analysis of variance was not feasible in view of the wide range of *S*s in each condition, 6–16. Individual *t* tests for normals yielded substantially the same significant comparisons as in the earlier analysis. The conditions of 1DR and 2DR were significantly superior to C ($p < .001$ and $< .05$, respectively) and in turn, C > 2D-IR ($p < .05$). The only significant exception to the predicted array of means was the mean trials of 1D-IR which was not significantly different from that of the relevant verbalization conditions. When we consider, however, that this mean score of 4.31 is based on only half of the original *S*s per cell and may represent relatively fast learners within the cell or reflect some other uncontrolled artifact, we are not surprised that a given score may go against prediction. For retardates, the range of mean scores was circumscribed and the only significant comparison was between 1DR and C ($p < .05$). In general, it may be concluded that fewer *S*s in the C condition solved the discrimination task than *S*s given relevant verbalization, and that those who did reach criterion in the C condition required more trials to do it.

TABLE 3
MEAN TRANSFORMED TOTAL TRIALS OF REVERSAL LEARNING IN NORMALS AND RETARDATES

Condition	Normal			Retardates		
	Subjects reaching criterion	*M*	*SD*	Subjects reaching criterion	*M*	*SD*
1DR	16	3.59	0.51	13	4.23	1.59
2DR	12	3.83	1.45	11	4.44	1.89
3DR	12	3.80	0.66	8	4.85	2.10
C	9	5.01	1.25	11	4.33	1.05
1D-IR	5	5.90	2.43	5	5.46	2.45
2D-IR	3	6.93	2.00			

Turning to reversal shift, frequencies and mean transformed total trials scores for both groups are presented in Table 3. It is observed that the majority of *S*s, 105 out of 120, were successful in solving the reversal shift. Of the 15 failures, only 5 were retarded, so that retardates did not contribute a disproportionate share. Of these 15 *S*s, 7 were in the entirely irrelevant verbalization conditions and the failure ratio was 7 out of 20 as compared to 8 out of 100 for the remaining conditions. A chi-square corrected for continuity yielded a value of 9.90 ($df = 1, p < .01$).

The array of mean reversal trials was in the predicted direction, but the range of scores was reduced especially for retardates where there were no significant differences between conditions. For the normals 1DR > C ($p < .01$) while all other comparisons were in the predicted direction, but did not attain formal significance. There were no overall significant differences between normals and retardates in reversal shift. A comparison of discrimination learning versus reversal shift trials also yielded no significant differences. As might be expected from the similar array of means scores by condition in discrimination learning and reversal shift, there was a highly significant product-moment correlation coefficient, .40 and .37 for normals and retardates, respectively (significant beyond .01).

DISCUSSION

Modifications in the two-choice discrimination-reversal paradigm were only partially successful in demonstrating the facilitative effect of relevant verbalization and interference due to irrelevant verbalization. With reference to discrimination learning, evidence is ample that verbalizing the relevant cue significantly improved performance over a base-line or no-verbalization condition and that verbalization of irrelevant cues interfered. Not only did more *S*s successfully reach criterion, but those who did, reached it more rapidly and the reverse was true when irrelevant cues were verbalized. This finding is consistent with the notion that verbal designations increase the probability that *S*s will attend to the stimulus parameters to which these designations refer. Those parameters that are consistently reinforced will be increasingly attended to, while the attending responses that are reinforced on a 50% or chance basis are gradually extinguished. This explanation agrees with such theoretical positions as House and Zeaman (1959) and Wyckoff (1952) who have formulated concepts of "observing" or "attending" responses and allow for a variety of variables, verbal labels, novelty, etc., to increase the probability of attending to a given class of cues.

When we consider the adverse effect of the consistent verbalization of irrelevant dimensions in 1D-IR and 2D-IR, we might well conjecture that the mixed condition, verbalizing two or more dimensions, one or more of which is irrelevant every time *S*s make a conceptual choice, should drastically reduce rate of learning even below that found in the control condition. This was not the case, and 2DR and 3DR were as effective, if not more effective, than the control condition and were certainly more effective than conditions in which verbalization was entirely irrelevant. If the net effect was not always beneficial when compared with this base-line condition, it was certainly not detrimental. We may conclude that *S*s recited irrelevant terms without necessarily attending to all the referent classes of cues and thereby being distracted from the criterial class. The *S*s did not attend equally to everything they said, but attended more to those verbal cues which were consistently rein-

forced and presumably ignored those dimensions which they uttered and which were inconsistently reinforced.

The data from reversal learning does not provide equally impressive evidence in support of the effects of verbalization. This is due partly to a marked reduction in the number of *S*s in reversal shift because of the many *S*s who failed to master the original discrimination task within the allotted trials and were not exposed to the reversal shift phase. Differences between means of the various conditions were generally in the predicted direction and with a larger number of *S*s per condition these differences might have attained formal significance at least in the case of the normal children.

The finding that overt relevant verbalization facilitates performance does not exclude the possibility raised by Kendler that covert verbalization may spontaneously occur in children age 7 or 8. We may assert, however, that providing *S*s with relevant verbal designations apparently directs their attention to criterial cues more rapidly than would occur when *S*s are left to their own resources.

The mechanism responsible for the deleterious effect of irrelevant verbalization in the reversal shift phase is not entirely clear. Generally, the more trials *S*s required to attain the correct discrimination, the more trials that were necessary to attain the reversal. Since this greater difficulty between the two phases was experienced by *S*s largely in the interfering conditions, we could argue that the fact of verbalization of irrelevant cues which continued throughout both phases served to strengthen erroneous attending responses. Another viewpoint would ignore the effect of the continuing irrelevant verbalization and would stress the failure of these *S*s to master the original discrimination as strongly as *S*s in the facilitating conditions. At the point of transition to reversal shift the former *S*s may have met the same experimentally determined criterion level without attaining thereby an implicit mediating response of comparable distinctiveness. It is planned to examine this question in a future study comparing reversal shift with and without continuing verbalizations of irrelevant cues.

There were no overall statistical differences between normals and retardates, but the predicted differences found greater confirmation for the normal children than for the retardates. This was especially true in the reversal shift phase where retardates performed alike regardless of verbalization, while significant effects of verbalization were still attained for the normal children. This suggestive finding is consistent with views of O'Connor and Hermelin (1959) and of Milgram and Furth (1967) that retardates are especially deficient when attempting to utilize overt verbal responses to regulate their behavior.

There are several implications for educational practice. Recitation aloud in the classroom may be beneficial, depending on what is being recited and what is being learned. If children are attempting to discover or identify the relevant cues in a complex stimulus, verbal responses referring to the relevant cues themselves will direct attention more immediately and exclusively to the requisite solution. Not only is the formal criterion of discrimination learning attained more rapidly, but the actual comprehension of the requisite solution is enhanced. If we assume that ease of reversibility of an acquired discrimination reflects superior conceptual or mediational efficacy, then the more rapid attainment of the reversal shift precisely by the children who had rapidly attained the original discrimination suggests that the original learning was of a flexible mediational character. Recital of relevant cues does not entirely eliminate the discovery aspect of learning, but may reduce sharply the generating of unclear and erroneous hypotheses. So advantageous is the directing of attention to the relevant cues that the children may apparently benefit from a complex verbalization which contains both facilitating and interfering verbal labels. Despite the "noise" in the channel, essential information apparently gets across, since a combination of relevant and irrelevant verbal cues yields superior performance to no verbalization at all.

One should not conclude, however, that

recital aloud is invariably helpful. In the present situation each instance of recital is accompanied by immediate feedback as to the correctness of the response made and under these circumstances it probably would direct exclusive attention to the criterial and away from the noncriterial cues.

REFERENCES

House, B., & Zeaman, D. Position discrimination and reversal in low-grade retardates. *Journal of Comparative and Physiological Psychology,* 1959, **52,** 564–565.

Kendler, T. S., & Kendler, H. H. Effect of verbalization on reversal shifts in children. *Science,* 1961, **134,** 1619–1620.

Kendler, T. S., Kendler, H. H., & Wells, D. Reversal and nonreversal shifts in nursery school children. *Journal of Comparative and Physiological Psychology,* 1960, **53,** 83–88.

Luria, A. R. The role of language in the formation of temporary connections. In R. Simon (Ed.), *Psychology in the Soviet Union.* Stanford: Stanford University Press, 1957. Pp. 115–129.

Milgram, N. A., & Furth, H. G. Factors affecting conceptual control in normal and retarded children. *Child Development,* 1967, **38,** 531–543.

O'Connor, N., & Hermelin, B. Discrimination and reversal learning in imbeciles. *Journal of Abnormal and Social Psychology,* 1959, **59,** 409–413.

Plenderleith, M. Discrimination learning and discrimination reversal learning in normal and feeble-minded children. *Journal of Genetic Psychology,* 1956, **88,** 107–112.

Reese, H. W. Verbal mediation as a function of age level. *Psychological Bulletin* 1962, **59,** 502–509.

Spiker, C. C. Verbal factors in the discrimination learning of children. *Monographs of the Society for Research in Child Development,* 1963, **28,** 53–69.

Stevenson, H. W. Discrimination learning. In N. R. Ellis (Ed.), *Handbook of mental deficiency.* New York: McGraw-Hill, 1963. Pp. 424–438.

Stevenson, H. W., & Zigler, E. F. Discrimination learning and rigidity in normal and feeble-minded individuals. *Journal of Personality,* 1957, **25,** 699–711.

Wyckoff, L. B., Jr. The role of observing responses in discrimination learning. Part I. *Psychological Review,* 1952, **59,** 409–413.

(Received April 18, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 3, 176–180

ATTITUDE LEARNING IN CHILDREN

C. JOAN EARLY
University of California, Berkeley

60 4th- and 5th-grade middle-class children were Ss in an experiment which was designed to demonstrate the effects of classical conditioning upon attitude behavior. Hypothesis 1 was supported by the significant changes which occurred in free play behavior in the case of each experimental isolate. Classmates approached experimental isolates more frequently than they did control isolates. The control isolates' behavior and interaction rate fluctuated insignificantly. The changed number of interactions remained at a consistent level after treatment for 1 wk. Hypothesis 2 was less directly supported because the sociogram data were less clearly changed. In one class, however, the high-popular children did change significantly. There were no significant changes in the low-popular halves of either class. Measurable changes were found after treatment which used a classical conditioning paradigm.

According to Doob (1947), an attitude is an implicit response which is both anticipatory and mediating in reference to patterns of overt responses. An attitude is evoked by a variety of stimulus patterns as a result of previous learning or of gradients of generalization and discrimination which is itself cue- and drive-producing. Using this definition, the present study involves three assumptions. (*a*) Individuals may be understood as discriminable stimulus patterns to which responses may be learned. (*b*) The name of any known person acts as part of the stimulus pattern which will recall the whole stimulus pattern to the subject's (*S*'s) memory. (*c*) Individuals in groups have acquired a kind of status value or stimulus value which makes them important or unimportant, friends or nonfriends of the other members of the group, varying with the others' experiences with them. The unimportant or low value person—the conditioned stimulus (CS) in this experiment—is defined as one who receives few social responses and who is regarded by no one or by one person only in his class as a friend (as measured by sociogram). He was called an isolate or social isolate for the purposes of this study.

Two hypotheses were investigated in the study. (*a*) It was anticipated that the attitude changes as a result of treatment would have an effect upon the actual behavior of the previously uninterested group members regarding the isolate, i.e., the treatment will mediate the overt responses of the experimental *S*s. (*b*) Attitude behavior which may be assessed by sociogram testing will change after using classical conditioning techniques to attach stimulus value to the isolate.

Staats and Staats (1958) associated pleasant and unpleasant meaning to words. He used these conditioned words to control the liking or disliking of new words when they were paired with the previous CS words. Like and dislike for the words in both levels of conditioning were measured by a rating scale. He further used national names (e.g., Dutch) as well as the names of people (e.g., Bill) as CSs. Depending upon the pleasantness of the unconditioned stimulus (UCS), the names were rated on a continuum from pleasant to unpleasant.

Conditioned behavioral changes in social approach behavior had not been noted before the current work. However, pilot studies by Early and Mercer (1961)[1] demonstrated that more children stated that they liked a social isolate or neutral stimulus after the name of that person had been paired with a positive evaluative meaning word (e.g., fun).

The current study was designed to note changes in sociogram ratings as well as be-

[1] Unpublished manuscript entitled "Attitude Change of a Group of Children by Classical Conditionings," 1961.

havioral changes after conditioning treatment. The individual treatment of each student was undertaken in order to determine the extent of his memory of stimulus pairs.

Method

Two classes of children in the fourth and fifth grades of the University of California laboratory summer elementary school were used as *S*s. They had attended class for 3½ weeks at the onset of the study. The sample was a middle-class group of children which was accumulated from districts in the Berkeley City Area. Of the 60 children, 95% were from families whose support was derived from fathers who were professional people. The teachers, observers, and first sociogram indicated six children who were alone most of the time during recess or free play periods. Preferred acquaintances had already been made. Groups of the more popular children were established in play activities.

In order to assess the group structure which existed in this temporary school setting, each teacher gave the following instructions for the first sociogram to his class.

> Please take a fresh sheet of paper and put your name on the upper right-hand corner. Think, without looking around, of some of the people you like in this classroom. Please write their names (first name and last initial if you know it) on the paper, one under the next. List as many as you like, but fewer than 10 or 11. We will probably be using this information for possible seating arrangements in the future.

A second sociogram similar to the first was administered after experimental treatment. This one was also allegedly for the teacher's use.

Using the first sociogram, each child was ranked according to the number of times he was mentioned by his classmates. The boys and girls in each class who received the fewest number of votes were designated social isolates. None of the isolates initially received more than one vote each. The most and least popular boys and girls in each class were divided randomly into experimental (E) and control (C) groups. This procedure developed eight groups: (*a*) 9 high-popular male experimentals; (*b*) 8 high-popular male controls; (*c*) 8 low-popular male experimentals; (*d*) 9 low-popular male controls; (*e*) 7 high-popular female experimentals; (*f*) 6 high-popular female controls; (*g*) 6 low-popular female experimentals; (*h*) 7 low-popular female controls. The total sample was 60 (30 in each class). The different numbers in each cell were a result of different numbers of boys and girls per class.

Two observers were used to note the behavior of each isolate for ½ hour each of 4 days: before experimental treatment, 1 day after treatment, 2 days after treatment, and 1 week after treatment. The number of interactions between the isolate and his school peers was counted. An interaction was counted for the isolate if he responded to or received a response from another child, that is, hugging, smiling, talking, or wrestling. Two C children (one boy and one girl having nearly the same number of votes and interactions as the isolates from their class) were observed in order to determine the effects of the mere passage of time. Without observing such C children, one could not be certain whether any of the isolates would "warm up" to a group of potential associates with repeated contact only.

The *S*s were given a list of words to rate on a 3-point scale: "I like," "I don't like," or "I don't care." Lists of high-positive evaluative meaning words and neutral words were developed from the children's responses to the rating scales. The most frequently liked words were used as UCSs. The words most frequently chosen by the girls as "I like" were the following: CONSIDERATE, GOOD, FUNNY, FRIENDLY, HAPPY, INTERESTING, PLAYFUL, SKILLFUL, KIND, POLITE, NEAT, NICE, GENEROUS, and CHEERFUL. The positive evaluative meaning words chosen by the boys were the same with the exception of POLITE and the addition of ACTIVE and FUN. The children were indifferent to the following words: AND, IF, OR, A, AN, FOR, OF, TABLE, and CHAIR

In order to prepare the children for experimental treatment the experimenter explained to the classes that she was conducting a study on memory. Each child would have a chance to read a list of paired words which would seem strange, then the children would be rated on their ability to recall as many pairs of words as possible. In general, the children were cooperative and industrious.

Each child was shown 32 cards after being taken individually into a room away from the class. Ten of the 32 cards in both E and C groups contained the name of the isolate. The other 22 cards contained names of class members paired with a low-valued word. One series of 32 cards read and recited (as a memory task) constituted one trial. The children were given individual, consecutive trials until a total of one-half of the conditioning pairs were learned and 70% of the conditioning pairs were memorized. In this way control over the extent of conditioning was established.

Experimental Group. The *S*s were shown the stimulus cards with the name of the isolate paired 10 times with one of the 10 positive meaning words. The boys were shown only boys' names, girls only girls' names. An example of the stimulus pairs containing a CS and a UCS follows:

> KAREN (isolate's name or CS)
> NEAT (positive evaluative meaning word or UCS)

Every other name (of the same-sexed class group)

appeared in the series paired with a nonevaluative word as in the following example.

MARY (nonisolate, classmate)
AND (nonevaluative word)

Control group. This randomly selected half of each class read the same number of words. However, they read the isolate's name paired with a nonevaluative word 10 times in the series. The control was used to note any changes in sociogram behavior which were not a result of experimental treatment.

KAREN (CS)
AND (nonevaluative meaning word or neutral stimulus)

Two forms of control were used, one for sociogram data and one for observational data. Reliability of social isolation could be established in a case where no treatment was given. If untreated isolates remained at the same level of interaction while the treated isolates changed, it was reasonable to conclude that experimental treatment was responsible for such observed differences.

ANALYSIS AND RESULTS

Two statistical tests were used in the data analysis: (*a*) the chi-square one-sample test for significance of change and (*b*) the Fisher Exact Test. The former test was used to examine the significance of change in the data obtained from behavioral observations of interactions made while the children were playing. The Fisher Exact Test was used to ascertain the significance of change between pretreatment and post-treatment sociograms. Results were considered significant if obtained p values were better than or equal to .05. The sociogram data were analyzed by class, sex, popularity, and E and C groups (Siegel, 1956).

The increment in number of observed social interactions after the initial observations was significant at the p equal to .01 significance level. For all treated isolates this increment was two to three times the initial rates. Table 1 contains raw data derived from observations of both experimental and control isolates. The increment of interactions in the two control cases was not significant. The first hypothesis which stated that there are behavioral changes following classical conditioning procedures was confirmed.

TABLE 1
NUMBER OF INTERACTIONS OBSERVED FOR EXPERIMENTAL AND CONTROL ISOLATES

Isolates	Before treatment	After treatment		
		Day 1	Day 2	1 Week
Experimental				
Female 1	9	24	32	30
Female 2	11	27	28	22
Subtotal	20	51	60	52
Experimental				
Male 1	8	31	26	27
Male 2	5	23	24	27
Subtotal	13	54	50	54
Control				
Female	10	10	12	5
Male	1	1	1	2
Subtotal	11	11	13	7

TABLE 2
INCREMENTS IN NUMBER OF VOTES FOR THE ISOLATE OF THE COMBINED HIGH-POPULAR MEMBERS OF THE FOURTH AND FIFTH GRADES

Isolates	No change in vote	Change in vote	Total *Ss*
Control	12	3	15
Experimental	8	7	15
Total *S*	21	9	30

The remaining data, regarding Hypothesis 2, are presented for the combined classes and Classes 1 and 2 which correspond to fourth and fifth grades. The number of votes by the high-popular children for the isolates was compared on Sociograms 1 and 2. The Fisher Exact Test was used to analyze these data. The second prediction, that classical conditioning would change the children's statements regarding whom they liked, was not confirmed.

Table 2 combines the two classes for the results of the votes from the high-popular groups. Changes in the low-popular half of the class were so low that no analysis was done. The changes for the combined boys' and girls' E groups for the high-popular half of the class were not significant.

Tables 3 and 4 contain the data from the fourth grade class. Table 3 shows the changes in the combined girls' and boys', high- and low-popular, E and C groups.

TABLE 3

INCREMENT OF VOTERS FOR ISOLATE AFTER TREATMENT, FOR FOURTH GRADE

Isolates	Boys E	Boys C	Girls E	Girls C	Total *Ss*
High Popularity	2	0	3	0	15
Low Popularity	0	1	0	0	15
Total Ss	8	8	7	7	30

These findings were not significant. Table 4 shows the changes for the high-popular E and C groups from Class 1. The increase in the number of votes from the first to the second sociogram was significant. The effects of treatment are clear here.

The second class contained 6 fourth-graders and 24 fifth-graders. None of these results was significant. The data in Table 5 illustrate that there were as many voters in the control group for the isolate as there were in the experimental group among the more popular children.

Conditioning of a few children appears to have been effective in both classes as measured by their vote changes for the isolate on the sociograms. However, the number of the children who were conditioned, as so measured, was significant in only one part of one class. Increments in the control group explained the insignificant results in the combined classes. In no case were the results of conditioning treatment in the low-popular group of *Ss* significant. Only 3 children of 60 voted for the isolate in this group, one of whom was in the control series.

The first hypothesis was strongly supported. The children did begin responding to the same-sexed isolate in their class after conditioning treatment. The second

TABLE 4

VOTING INCREMENTS IN THE HIGH-POPULAR EXPERIMENTAL AND CONTROL GROUPS FOR FOURTH GRADE

Isolates	No change in vote	Change in vote	Total *Ss*
Control	8	0	8
Experimental	2	5	7
Total *Ss*	10	5	15

TABLE 5

INCREMENT OF VOTERS FOR ISOLATE AFTER TREATMENT, FOURTH AND FIFTH GRADES COMBINED

Isolates	Boys E	Boys C	Girls E	Girls C	Total *Ss*
High Popularity	1	0	2	3	15
Low Popularity	1	1	0	0	15
Total *Ss*	9	9	6	6	30

hypothesis was clearly supported by the data from high-popular children in one class only. Sociogram behavior was not changed significantly as was actual play behavior.

There were no significant differences in boys' increments between Sociograms 1 and 2. The E girls' changes did not achieve significance over the C girls' group in either combined classes or separate classes.

DISCUSSION

Initial observations of the isolated children revealed that they tended to be unresponsive to others. They did very little during free play at recess and only infrequently spoke to other children. If they spoke at all it was to someone smaller or younger than themselves. During the period of treatment and observations there was a noticeable change in the quality of their behavior. All four of the isolates became more animated, participated in games with other children when they had not done so previously, and chose children from their own class as playmates. In one case, immediately after treatment, a boy isolate left the younger children's game and joined his own class for the first time, where he stayed during the remainder of the observation periods. Conditioning treatment of his peers produced responses to his social overtures so that they noticed him and began to offer themselves as social reinforcers. Because the isolate's peers reinforced his approach behavior, it was continued when extinction might have occurred without reinforcement.

Qualitative data seem to explain the findings from the second class pertaining to

Hypothesis 2. In the process of arranging a random, stratified sample, the peer groups of the high-popular children were broken into experimental and control groups. This artificial division did not reduce the group preferences which were established before the experiment. Thus, a treated child approached the isolate accompanied by his friends who were in the C group. Probably because a few of the experimental popular children associated with the social isolate, the control children learned to value the isolate and therefore voted for him as a person whom they liked. Nonexperimental conditioning seemed to have occurred. The popularity of one child was associated with the isolate, thus probably making him acceptable to other high-popular children.

The method of selecting the controls, then, made the experimental effects clear as well as the effects of indirect conditioning. In the second class, the experimenter conditioned three children, and the treated children conditioned as many. The results from the second class were not significant at p better than or equal to .05; however, the first class' data supported the second hypothesis.

The low-popular children included untreated isolates who were quiet, inactive, and unresponsive. In general they reported fewer children whom they liked. Many of these *S*s actually avoided interaction. They required more trials to the learning criterion. Further work in this area might demonstrate some personality differences in resistance, passivity, defensiveness, or introversion. On the other hand, the high- and medium-popular children seemed to be more willing to participate in class discussions, games, and memory tasks. They were more socially active, hence more receptive and responsive to stimuli outside of themselves.

A previous, unpublished study by Early and Mercer (see Footnote 1) demonstrated that high-popular children, particularly girls, would be more likely to state that they liked the isolate after conditioning treatment than would high-popular boys. The present study also found that the raw score changes for girls were greater than for the boys despite the lower number of girls. It seemed that girls were more conditionable or suggestible than the boys because of their greater changes after treatment. This tendency was not significant, however.

Further work in the area might include assessing personality differences between high- and low-popular children. IQ differentials may account for the decreased ability to memorize lists of pairs. If the less bright fall into the low-popular group, they, therefore, may seem less attractive to the brighter, more verbal, popular children. Other personality reasons for a child's tendency to remain socially isolated might include shyness, fear, or disinterest in the world outside of themselves. The last case implies that the genuinely introverted child has chosen to remain unnoticed.

REFERENCES

DOOB, L. W. The behavior of attitudes. *Psychological Review,* 1947, **54**, 135–156.

SIEGEL, S. *Nonparametric statistics.* New York: McGraw-Hill, 1956.

STAATS, A. W., & STAATS, C. K. Attitudes established by classical conditioning. *Journal of Abnormal and Social Psychology,* 1958, **57**, 37-40.

(Received April 25, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 3, 181–185

ACHIEVEMENT MOTIVATION AND THE RECALL OF INCOMPLETED AND COMPLETED EXAM QUESTIONS[1]

BERNARD WEINER, PATRICK B. JOHNSON, AND ALBERT MEHRABIAN
University of California, Los Angeles

Male students indicated their level of aspiration on a final exam, and subsequently were asked to recall the exam items. 2 measures of resultant achievement motivation, 1 objective and 1 in part projective, were employed to classify Ss into motive groups. For both measures Ss high in resultant achievement motivation recalled a greater percentage of failed than passed questions (the Zeigarnik effect). In addition, they exhibited a greater Zeigarnik effect than Ss low in resultant achievement motivation. The differential recall was due to greater remembrance of the failed items by the high achievement-oriented Ss. It is hypothesized that these students covertly rehearse and think about the missed questions more than students low in achievement motivation. Therefore, it is contended that the Zeigarnik effect is a learning rather than a memory phenomenon. Only the projective measure revealed group differences in level of aspiration.

The interrupted task paradigm was introduced into psychology by Zeigarnik in 1927. Individuals receive a number of tasks to complete, and are interrupted before finishing some of them. Following this activity, they unexpectedly are asked to recall the tasks. Zeigarnik found greater recall of the incompleted (I) than completed (C) tasks (the Zeigarnik effect). The differential recall was believed to support Lewin's (1935) conception of enduring tension systems.

There was a partial reversal of Zeigarnik's results when studies of task recall were conducted in America. Many investigators (e.g., Glixman, 1949; Rosenzweig, 1943) found greater recall of the C than I tasks in "ego-involved" situations. It was reasoned that it is "threatening" to remember failure (I) experiences; the material associated with failure therefore is "repressed."

Atkinson (1953) in part resolved the apparent contradiction between the findings of Zeigarnik and Rosenzweig. He demonstrated that individuals classified as high in need for achievement (n Ach) remember more I than C tasks in achievement-oriented contexts. Conversely, subjects (*S*s) who are low in n Ach and considered relatively anxious about failure remember more C than I tasks. Analysis of *S*s used by Rosenzweig indicated that they were receiving services from the psychological clinic, and presumably would be characterized as relatively anxious. This was not true of the population used by Zeigarnik. Atkinson argues that the different *S* populations were responsible for the contradictory results of Zeigarnik and Rosenzweig. Atkinson also found that the differential Zeigarnik effect was attributable to differences in the recall of I, rather than C, tasks. He reasons that the remembrance of I tasks is instrumental to the attainment of achievement-related goals; these goals are strived for by individuals high in n Ach, but avoided by individuals relatively low in n Ach. (The reader is directed to Butterfield, 1964, and Weiner, 1966a, for a more detailed review of experimentation in this area.)

In the present study Atkinson's suppositions are investigated in a real-life achievement setting. Following a final examination students were asked to recall the exam items. These circumstances should maximize aroused achievement motivation and task involvement, and therefore mag-

[1] This research was supported in whole by Public Service Research Grant No. MH-12603-03 from the National Institute of Health awarded to the senior author. The authors wish to thank Paul Feldman, Carol Price, and Penelope Potepan for their aid in conducting the experiment.

nify previous findings (Atkinson, 1953; Zeigarnik, 1927). Two measures of *resultant* achievement motivation (n Ach minus anxiety) were employed to classify Ss into motive groups. One classification method included a Thematic Apperception Test (TAT) to measure n Ach, and a Test Anxiety Questionnaire (TAQ) to assess level of anxiety. Subtracting the z score on the TAQ from the z score on the TAT yields a measure of resultant achievement motivation. This method of grouping individuals is used most extensively in current studies of achievement motivation (Atkinson, 1964). The second measure of resultant achievement motivation was devised by Mehrabian (in press), and is similar in principle to a self-report measure constructed by O'Connor (1962). The O'Connor scale has been used with some success (e.g., Weiner, 1966b) to classify Ss into motive groups.

METHOD

The Ss were 205 students enrolled in the undergraduate personality course at the University of California, Los Angeles. On the second day of class a TAT, picture series 2, 8, 4, 48 (Atkinson, 1958) was administered under neutral conditions (McClelland, Atkinson, Clark, & Lowell, 1953). All pictures were highly cued for achievement. The story protocols were scored for n Ach by a trained rater according to a reliable method of content analysis (Atkinson, 1958).[2] Evidence suggests that the TAT measure of n Ach is not a valid indicator of achievement strivings for all females (Atkinson, 1958; French & Lesser, 1964). Only the male sample ($N = 82$) was used in the final data analysis. Following the administration of the TAT, the TAQ was distributed (Mandler & Sarason, 1952). This is a self-report measure of situationally aroused anxiety. The items were scored on a 5-point Likert Scale. As previously indicated, z scores on the TAT and TAQ were computed, and an index of resultant achievement motivation derived by subtracting z score on the TAQ from the z score on the TAT. The Ss in the top and bottom 25% of the distribution were respectively classified as high or low in resultant achievement motivation, while the remaining 50% of the sample comprised the middle group.

The final test administered was a measure of resultant achievement motivation devised by Mehrabian (in press). The test, labeled the Resultant Achievement Motivation Scale (RAM), includes 34 items primarily derived from a theory of achievement motivation formulated by Atkinson in 1957 and from data supporting that conception. Atkinson and other investigators have demonstrated that individuals high in resultant achievement motivation engage in achievement-related activities, anticipate success, and prefer tasks of intermediate difficulty. The test items tap the direction of behavior, approach or avoidance, exhibited in achievement contexts; the kind of affect, hope or fear, associated with achievement situations; and the degree or risk, intermediate versus easy or difficult odds, preferred. Sample items are:

1. In my spare time I would rather learn a game to develop skill than for recreation.
2. I worry more about getting a good grade than I worry about getting a bad grade.
3. I would prefer a job which is important, difficult, and involves a 50% chance of failure, to a job which is somewhat important but not difficult.

Items were rated +3 (very strong agreement) to −3 (very strong disagreement). The 10-week test-retest reliability of the measure is $r = .78$; the item-total correlations range between .2–.5.[3] Again Ss high or low in resultant achievement motivation comprised the top and bottom 25% of the distribution.

Ten weeks after the individual difference test administrations, the students were given a 58-question final examination. At the top of the test the following question was written:

I am trying to get ______ correct out of 58.

The responses to this question provide an index of level of aspiration on the exam (Lewin, Dembo, Festinger, & Sears, 1943). The test format was "fill in the blanks." When the students handed in their exams, they were given a paper with the following written instructions:

> Please start at the top of the space below and write, as they occur to you, the items on the exam. Do not worry about spelling or how exact your memory of each question is. You do not have to write the whole question, but just enough for us to be able to identify it. The questions need not be written in the order in which they appeared on the exam. Look at the clock and take three minutes to do this. At the end of that time turn in the paper.

To provide an objective index of task completion, test items missed were considered incomplete or failed, while correct items were considered complete or successful. Some failed items un-

[2] The protocols were scored by Patrick Johnson. Interrater reliability with the senior author had been established to be $r = .86$.

[3] Further details of test construction, reliability, and validity will be presented in a forthcoming paper by Mehrabian.

doubtedly were subjective successes, and some solved questions subjective failures. These occurrences should attenuate the expected results (Marrow, 1938). Because of the nature of the examination situation, it was decided not to ask the students to indicate their perceived performance on each question during the exam.

RESULTS

The correlation between the two indexes of resultant achievement motivation was positive and significant, but relatively low, $r = .30$, $p < .01$. The mean number of correct answers on the exam was 46. Therefore, there were many more C than I items. Exam performance was virtually identical for the three motive groups when classified with either resultant motivation index ($F < 1$).

Table 1 gives the percentage of I–percentage of C recall and percentage of *S*s exhibiting a Zeigarnik effect for the three motive groups. The table indicates that the likelihood of a Zeigarnik effect is monotonically related to the strength of resultant achievement motivation. This occurs when either the TAT-TAQ or RAM is used as the motive measure. Among *S*s high on the TAT-TAQ index, 15 out of 20 ($p < .05$) exhibit a Zeigarnik effect, while of the 21 *S*s high on the RAM, 17 recall a greater percentage of I than C items ($p < .01$). The difference between the proportion of *S*s in the extreme groups recalling a greater percentage of I than of C questions approaches statistical significance with the TAT-TAQ index, ($z = 1.67$, $p < .10$), and is significant when the RAM is the motive measure ($z = 2.26$, $p < .05$). The prediction for the differential recall is best substantiated when both indexes are employed simultaneously to classify *S*s into motive groups. Of the 9 *S*s high on both measures, 8 show a Zeigarnik effect; only 2 of the 6 *S*s low on both resultant measures recall a greater percentage of I than C test items ($p < .05$, Fisher Exact Test).

TABLE 1

RECALL OF INCORRECT MINUS CORRECT ITEMS AND LEVEL OF ASPIRATION FOR GROUPS DIFFERING IN STRENGTH OF RESULTANT ACHIEVEMENT MOTIVATION

Item	Motive Measure					
	TAT-TAQ			RAM		
	High	Mid	Low	High	Mid	Low
N	20	42	20	21	40	21
Percentage of *S*s with percentage of I > percentage of C recall	75	57	50	81	60	48
Percentage of I–percentage of C recall	10.0	5.3	−2.3	8.8	6.5	−2.8
Aspiration level	54.9	51.9	48.9	51.1	51.7	52.6

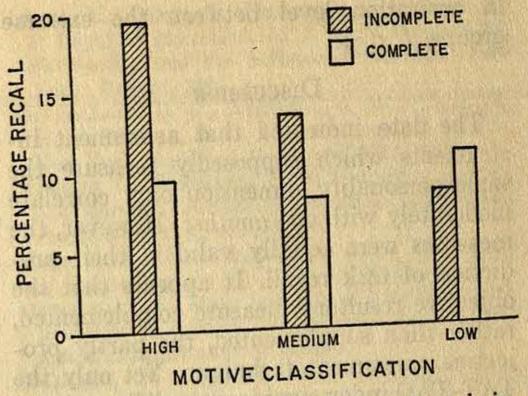


FIG. 1. Percentage recall of correct and incorrect items when *S*s are classified according to strength of resultant achievement motivation (TAT-TAQ).

Figure 1 separately illustrates the percentage recall of the I and C questions for the motive groups when classified according to TAT-TAQ score. The figure reveals clearly that the Zeigarnik effect shown in Table 1 is primarily due to the differential recall of I items. The *S*s high in resultant achievement motivation recall a greater percentage of I questions than *S*s low in resultant achievement motivation ($z = 1.82$, $p < .10$). The pattern of results is identical when *S*s are classified according to score on the RAM ($z = 1.68$, $p < .10$).

Level of Aspiration

Table 1 also shows the level of aspiration for the three motive groups. When the TAT-TAQ is employed to classify *S*s, there is a monotonic relationship between aspiration level and strength of resultant achievement motivation. The difference in aspiration level between the extreme motive groups is significant, $t = 2.35$, $df = 38$, $p < .01$. However, when the RAM is the motive measure, there are no differences

in aspiration level between the extreme groups, $t < 1$.

DISCUSSION

The data indicated that assessment instruments which supposedly measure the same personality dimension only correlate moderately with one another. However, the measures were equally valid in their prediction of task recall. It appears that the objective resultant measure complemented, rather than supplemented, the partly projective assessment technique. Yet only the TAT-TAQ index significantly differentiated the motive groups in terms of aspiration level. Level of aspiration in the present context denoted an imaginative achievement goal. Many students indicated that their goal was to answer all the questions correctly. The responses were similar to those of *S*s asked what they are "hoping" for, as opposed to what level they actually are attempting to reach (Lewin et al., 1943). Recently, Wallace (1966) has argued that: "the closer the approximation of the role-playing situation to the predictive situation, the greater... the accuracy of the prediction [p. 136]." In the present experiment the partial fantasy measure predicted fantasy behavior better than the objective index. It is conceivable that the objective and fantasy indexes of resultant achievement motivation will, at times, successfully predict different achievement behaviors. In sum the data presented here do provide validity for the RAM as a motive measure, and suggest that this is a promising instrument for the prediction of some achievement strivings. In addition, because the RAM items were derived from a theory of achievement motivation, the positive results tend to validate both the theory and the measure (Cronbach & Meehl, 1955).

The findings concerning task recall replicate results reported previously by Atkinson (1953). The Zeigarnik effect is predominantly manifested in *S*s high in resultant achievement motivation, and is caused by the differential recall of I tasks. But differential recall does not necessarily reflect memory disparities. Retention is conceptualized as a multistage process. The first stage is learning, or trace formation. The temporally subsequent stages are trace storage and trace retrieval. Only the latter two stages are adjudged to be memory processes. To show differences in the memory of events, there must be equality in the degree of original learning of the material. If *S*s learn I tasks to a greater degree than C tasks, then one also would expect differential recall of those tasks. Caron and Wallach (1957) have demonstrated that the Zeigarnik effect is due to differential learning rather than to differential retention. In their study *S*s were told that the I tasks were insoluble after the initial recall period was completed. Therefore, there was no persisting source of motivation for relatively anxious *S*s to repress the tasks, nor any instrumental inducement for *S*s striving for success to retain the material. Following this information about the insolubility of the I tasks the differential retention found at the end of the first recall period should dissipate. However, customary differences in the pattern of recall between the motive groups were observed after the feedback. Caron and Wallach therefore concluded that the recall differences must be attributed to differential learning, rather than to differential memory.

The results of the study by Caron and Wallach, combined with the present data, suggest that *S*s high in resultant achievement motivation learn I tasks to a greater extent than *S*s low in resultant achievement motivation. Learning is in part a function of the number of repetitions of the stimulus. It is hypothesized that *S*s high in resultant achievement motivation covertly repeat and rehearse questions which they have missed more than low-achievement individuals. Weiner (1965) has summarized a number of studies which reveal that *S*s high in achievement motivation are attracted toward tasks which they have initially failed. Conversely, *S*s low in achievement motivation especially avoid tasks which they have not been able to complete. Because students low in resultant achievement motivation avoid failed or incompleted test items, they are less likely covertly to repeat and remember those items than the high resultant achievement motivation students. This analysis implies that if students

are not allowed to return to the failed tasks, then the differential Zeigarnik effect would not be exhibited. The differential persistence at failed tasks also might be responsible for the disparate grade point averages that are, at times, manifested by the groups (see McClelland, et al., 1953).

REFERENCES

ATKINSON, J. W. The achievement motive and recall of interrupted and completed tasks. *Journal of Experimental Psychology*, 1953, **46**, 381–390.

ATKINSON, J. W. Motivational determinants of risk-taking behavior. *Psychological Review*, 1957, **64**, 359–372.

ATKINSON, J. W. (Ed.) *Motives in fantasy, action, and society*. Princeton: Van Nostrand, 1958.

ATKINSON, J. W. *An introduction to motivation*. Princeton: Van Nostrand, 1964.

BUTTERFIELD, E. C. The interruption of tasks: Methodological, factual, and theoretical issues. *Psychological Bulletin*, 1964, **62**, 309–322.

CARON, A. J., & WALLACH, M. A. Personality determinants of repressive and obsessive reactions to failure stress. *Journal of Abnormal and Social Psychology*, 1957, **55**, 372–381.

CRONBACH, L. J., & MEEHL, P. E. Construct validity in psychological test. *Psychological Bulletin*, 1955, **52**, 281–302.

FRENCH, E. G., & LESSER, G. S. Some characteristics of the achievement motive in women. *Journal of Abnormal and Social Psychology*, 1964, **68**, 119–128.

GLIXMAN, A. F. Recall of completed and incompleted activities under varying degrees of stress. *Journal of Experimental Psychology*, 1949, **39**, 281–295.

LEWIN, K. *A dynamic theory of personality*. New York: McGraw-Hill, 1935.

LEWIN, K., DEMBO, T., FESTINGER, L., & SEARS, P. S. Level of aspiration. In J. McV. Hunt (Ed.), *Personality and the behavioral disorders*. New York: Ronald, 1943.

MCCLELLAND, D. C., ATKINSON, J. W., CLARK, R. W., & LOWELL, E. L. *The achievement motive*. New York: Appleton-Century-Crofts, 1953.

MANDLER, G., & SARASON, S. B. A study of anxiety and learning. *Journal of Abnormal and Social Psychology*, 1952, **47**, 166–173.

MARROW, A. J. Goal tension and recall. II. *Journal of General Psychology*, 1938, **19**, 37–64.

MEHRABIAN, A. Male and female scales of the tendency to achieve. *Educational and Psychological Measurement*, in press.

O'CONNOR, P. An achievement risk preference scale: A preliminary report. *American Psychologist*, 1962, **17**, 317 (Abstract).

ROSENZWEIG, S. An experimental study of "repression" with special reference to need-persistive and ego-defensive reactions to frustration. *Journal of Experimental Psychology*, 1943, **32**, 64–74.

WALLACE, J. An abilities conception of personality: Some implications for personality measurement. *American Psychologist*, 1966, **21**, 132–138.

WEINER, B. The effects of unsatisfied achievement motivation on persistence and subsequent performance. *Journal of Personality*, 1965, **33**, 428–442.

WEINER, B. Effects of motivation on the availability and retrieval of memory traces. *Psychological Bulletin*, 1966, **65**, 24–37. (a)

WEINER, B. Achievement motivation and task recall in competitive situations. *Journal of Personality and Social Psychology*, 1966, **3**, 693–696. (b)

ZEIGARNIK, B. Das Behalten erledigter und unerlidigter Handlungen. *Psychologische Forschung*, 1927, **9**, 1–85.

(Received April 27, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 3, 186–190

VERBAL ORGANIZATION AND THE FACILITATION OF SERIAL LEARNING[1]

JOEL R. LEVIN AND WILLIAM D. ROHWER, JR.
University of California, Berkeley

96 4th- and 5th-grade children participated in a serial learning experiment. Ss were required to learn a 14-item serial list in either the traditional manner, or with the aid of individual phrases, or in the context of a single sentence. It was found that Ss who were given the single unifying structure, the sentence, performed significantly better than did Ss who learned under nonunified conditions. These results contradict those of an earlier experiment in which it was concluded that verbal organization does not facilitate performance in a serial task.

Recently, considerable attention has focused on the question whether or not identical processes are involved in learning the serial order of items in a list and in learning the pair-wise arrangement of items in a list. The most frequently used method in attempts to answer this question has been that of transfer designs, either serial (Ser) to paired associate (PA) or PA to Ser. Although evidence thus far produced by the application of this method is not yet entirely conclusive, some investigators have construed available results as implying that the processes of Ser and PA learning do indeed differ. Jensen (1962), for example, has contended that the learning of a Ser list consists of the integration of a sequence of responses into a single unit rather than of the acquisition of connections between successive eliciting stimuli and their companion responses. Adopting this contention (Jensen & Rohwer, 1965a), Jensen and Rohwer (1965b) have gone on to characterize one of the differences between Ser and PA learning in terms of the relative importance of past verbal experience for the two kinds of learning:

> In short, we hypothesize that PA learning ability reflects relatively more the richness of *S*'s past verbal experience and its spontaneous availability in a learning situation, while serial learning constitutes a more fundamental kind of ability which is relatively unaffected by the amount of previous verbal experience [Jensen & Rohwer, 1965b, p. 602].

The purpose of the present experiment is to disentangle the hypothesis of response integration as a description of Ser learning and the assertion that the availability of previous verbal experience is irrelevant to the efficiency of Ser learning. The validity of the latter hypothesis depends, in part, on the results of a study (Jensen & Rohwer, 1965b) conducted to test one of its implications, namely, that verbal organization of PA items should facilitate acquisition whereas verbal organization of Ser items should not. The tasks of learning a 10-pair PA list and a 10-item Ser list were given to children of a variety of grade levels. In the case of both tasks, the treatment and control conditions were distinguished only by the procedure followed on the initial study trial. For the latter, the PAs and the Ser items were simply shown successively and *S* was asked to name the object in each picture as it was shown to him. In the treatment condition for the PA task, *S* was asked to elaborate the names of the two objects in each pair into a sentence, one sentence per pair. Similarly, the treatment condition for the Ser task required *S* to elaborate the names of each successive pair of objects into a sentence, two items per sentence. Note that this procedure is consistent with the conception that Ser learning consists of the acquisition of con-

[1] This work was supported, in part, by a contract with the United States Office of Education (OE6-10-273) through Project Literacy. The report was prepared at the Institute of Human Learning, which is supported by grants from the National Science Foundation and the National Institutes of Health.

nections between successive items each of which serves both as a stimulus and as a response.

The results for elementary school children replicated those previously obtained for mentally retarded adults (Jensen & Rohwer, 1963): The sentence condition produced substantial facilitation of PA but not of Ser learning. Accordingly, the investigators concluded in favor of their original hypothesis regarding the irrelevance of verbal organization for Ser learning and went on to say:

> If a true difference between the sentence and naming conditions were found to exist, we would be inclined to interpret the difference as being attributable to facilitation of response learning rather than to facilitation of serial learning per se [p. 606].

In contrast to these conclusions, the guiding hypothesis for the present study is that verbal organization is relevant to Ser learning but only if the type of organization imposed is consistent with what is ordinarily learned when a Ser list is acquired. In agreement with Jensen (1962) and with Jensen and Rohwer (1965a) it is assumed that the learning of a Ser list consists of the process of integrating the items into a single response. On this assumption, the absence of facilitation previously reported (Jensen & Rohwer, 1963, 1965b) would be expected; the kind of verbal organization used, that is, successive discrete sentences, is not consistent with what is presumably acquired in Ser learning. A different type of verbal organization, specifically, one that confers on all items in the Ser list membership in a single unit, would be expected to produce facilitation not accountable in terms of enhanced response learning. The present experiment was designed to test this prediction.

METHOD

Design and Materials

All *Ss* were given a common task, namely, to learn the serial order of one or the other of two lists of 14 familiar nouns. The design was a 6 × 2 × 2 × 4 factorial in which the factors were, respectively: conditions, lists, grades, and trials. The various conditions differed only with regard to the character of the one study trial during which the list was first presented; thereafter all were identical.

The six conditions may be viewed as an aggregation of an experimental and five control groups. The study-trial materials for the experimental or single sentence (SS) condition were constructed to conform with the requirement that all of the items in the list be contained within the same verbal unit. Each of the 14 nouns was presented in the context of a three- or four-word phrase. The critical property of the phrases was that when read in the prescribed order, they formed one continuous, meaningful sentence. Thus the study-trial materials for the SS condition were conceived to be a concrete expression of a verbal organization consistent with the interpretation of serial learning as a process of integrating a single response.

In order to answer the question of central interest, that is, whether or not SS would facilitate Ser learning, a noun control (NC) condition was used. The study-trial materials for NC simply consisted of the 14 nouns in the list presented successively in accord with the traditional Ser procedure.

Since it was expected that learning would be more efficient in SS than in NC, an additional condition was necessary to permit an evaluation of an alternative interpretation of whatever facilitation was observed. Following Jensen and Rohwer (1965b) it might be argued that the verbal context in SS would affect response learning rather than the integration of the list as a unit. If so, it can be reasoned that the same effect would be produced by the presentation of each noun in the context of a phrase even when the phrases are independent of one another such that their seriation does not form a continuous sentence. Accordingly, in the phrase control (PC) condition, the study-trial materials consisted of a set of 14 unrelated phrases, one for each of the nouns in the list. The PC condition served the added function of providing a comparison with SS in which noun study time was equated, as was not the case with NC.

As a control for response learning, however, PC by itself was not entirely adequate since the particular words used in the verbal contexts were necessarily different than those used in SS. By way of obviating this difficulty, the same phrases used in SS were presented in a scrambled order in the scrambled sentence control (SSC) condition such that their succession did not form a sentence.

Note that although all of the words presented in SSC were identical with those in SS, the order of items in the list varied. Thus in order to evaluate directly the effects of SSC on serial learning, the remaining conditions in the design were simply scrambled versions of NC (SNC) and PC (SPC) where the order of the items in the list was the same as that in SSC. A complete set of the study-trial materials used for one of the lists is shown in Table 1.

TABLE 1
LIST A MATERIALS

Single sentence (SS)	Phrase control (PC)	Noun control (NC)
the grey CAT	the grey CAT	CAT
jumped over the LOG	we jumped the LOG	LOG
and crossed the STREET	I crossed the STREET	STREET
to find the BOWL	you find the BOWL	BOWL
of cold MILK	some cold MILK	MILK
under the CHAIR	his own CHAIR	CHAIR
in the new HOUSE	our nice new HOUSE	HOUSE
by the blue LAKE	a little blue LAKE	LAKE
where the young BOY	my fine young BOY	BOY
lost his left SHOE	he lost his SHOE	SHOE
while eating the FISH	she's eating the FISH	FISH
on the wooden BOAT	an old wooden BOAT	BOAT
during the STORM	that awful STORM	STORM
that came last YEAR	they came last YEAR	YEAR
Scrambled sentence control (SSC)	**Scrambled phrase control (SPC)**	**Scrambled noun control (SNC)**
that came last YEAR	they came last YEAR	YEAR
the grey CAT	the grey CAT	CAT
of cold MILK	some cold MILK	MILK
where the young BOY	my fine young BOY	BOY
and crossed the STREET	I crossed the STREET	STREET
on the wooden BOAT	an old wooden BOAT	BOAT
lost his left SHOE	he lost his SHOE	SHOE
to find the BOWL	you find the BOWL	BOWL
while eating the FISH	she's eating the FISH	FISH
jumped over the LOG	we jumped the LOG	LOG
in the new HOUSE	our nice new HOUSE	HOUSE
under the CHAIR	his own CHAIR	CHAIR
during the STORM	that awful STORM	STORM
by the blue LAKE	a little blue LAKE	LAKE

The three remaining factors were: lists, grades, and trials. Two distinct lists of nouns were used to reduce the risk that results would be specific to one set of items. Children were drawn from two grade levels rather than one only to provide a sample of adequate size, not to test hypotheses as to age differences.

Procedure

When *S* entered the room, the experimenter (*E*) told him that he was to memorize a list of nouns (or nouns in phrases) in the order in which they were presented. The instructions described the procedures that would be followed in the study trial and in the anticipation trials as well as the type and timing of the responses expected.

All materials were presented on a memory drum. Immediately after the instructions, the 14 successive nouns (or phrases) were shown at a 4-second rate and, as each one appeared, it was read aloud by *E*. Following the study trial an asterisk appeared and *S* had 4 seconds to supply the first noun. The first noun then appeared, and *S* had another 4 seconds to offer the second noun, and so on through the list to the end of the first anticipation trial. Three more anticipation trials were given for a total of four in all.

It is important to note that in all conditions, only the nouns themselves were presented during the four anticipation trials. In other words, *S*s in the sentence and phrase conditions were given a verbal context only on the initial presentation trial.

Subjects

Ninety-six fourth- and fifth-grade children from a school serving a middle class residential area participated in the experiment.[2] Forty-eight children from each grade were randomly assigned to the six experimental conditions. All *S*s were tested inidividually by *E*.

RESULTS

The dependent variable in this study was the number of correct responses given by *S* over the four anticipation trials. A repeated measures analysis of variance was performed on the data. The analysis of variance table is presented in Table 2. All hypotheses were tested with the probability of a Type I error equal to .05. It may be seen from Table 2 that there are three significant sources of variation, namely conditions, trials, and grades × trials.

The trials effect was expected, and accounts for about 57% of the within variance. The grades × trials effect may be traced to the slightly superior learning rate of *S*s in the fifth grade.

It is of particular interest that there is no main effect for either grades or lists, and that none of the interactions involving these factors in the between portion of the table is significant. The mean number of correct responses per trial as a function of conditions and lists is presented in Table 3.

Within the main effect of conditions, Scheffé's method for post hoc comparisons reveals that the SS group differs from each of the other groups, and that no other pairwise contrasts are significant. The com-

[2] The authors would like to express their appreciation to Elmer Venter and his staff at John Muir Elementary School for their cooperation in the execution of the present study.

TABLE 2

ANALYSIS OF VARIANCE TABLE

Source	*df*	*MS*	*F*
Between subjects	95	20.30	
Grades	1	6.51	—
Conditions	5	124.06	8.28*
Lists	1	20.17	1.35
G × C	5	9.91	—
G × L	1	1.04	—
C × L	5	24.42	1.63
G × C × L	5	6.13	—
Error	72	14.98	
Within subjects	288	5.06	
Trials	3	275.91	128.33*
G × T	3	6.61	3.07*
C × T	15	2.73	1.27
L × T	3	5.42	2.52
G × C × T	15	1.57	—
G × L × T	3	1.67	—
C × L × T	15	1.30	—
G × C × L × T	15	2.57	1.20
Error	216	2.15	
Total	383	8.84	

* $p \leq .05$.

parison involving SS versus the average of all other conditions combined accounts for 91.5% of the total between conditions sum of squares, and the sum of squares for all other available orthogonal comparisons is not significant. As an inspection of Table 3 indicates, the results are clear; SS did facilitate learning relative to the ordinary serial procedure condition, NC, and the magnitude of facilitation was as great on Trial 4 as it was across trials. The additional fact that NC produced as many correct responses as each of the other control conditions contraindicates an interpretation of the facilitory effect of SS in terms of an enhancement of response learning.

TABLE 3

MEAN NUMBERS OF CORRECT RESPONSES ACROSS TRIALS AND ON TRIAL 4

Lists	Conditions						
	SS	NC	PC	SSC	SNC	SPC	All
A	7.62	5.06	4.19	5.25	5.62	3.91	5.28
B	7.90	5.19	4.38	2.84	4.03	4.56	4.82
Across trials	7.76	5.12	4.28	4.05	4.83	4.23	5.05
Trial 4	10.12	7.25	6.00	5.75	6.50	5.44	6.84

DISCUSSION

The results of the present study support the initiating hypothesis, namely that verbal organization is relevant to the integration of a sequentially ordered set of responses. Even though this conclusion is in direct opposition to that reached by Jensen and Rohwer (1965b) it is not inconsistent with their interpretation of serial learning as a process of response integration (Jensen & Rohwer, 1965a). Indeed, the present results may be construed as indirect evidence in support of that interpretation since the form of verbal organization employed follows from it.

A fruitful theory of what is learned in serial learning ought to have implications for the design of conditions to facilitate that process. The adequacy of the theory, then, depends, in part, upon whether or not the facilitative procedures that can be derived from it serve to increase learning efficiency. Although the present results are suggestive, they are not sufficient to permit a conclusive judgment in this regard. Accordingly, it is of some import to conduct a comparative experiment designed to assess the relative efficacy of facilitative conditions derived from the principal theories of serial learning.

The problem of the effect of SS on response learning deserves brief additional comment. In the present design, no provision was made for a direct assessment of the degree of response learning as a function of study-trial conditions. Nevertheless, it is difficult to discern in the SS phrases any properties relevant to the efficiency of response learning that are not also present in the PC phrases. Thus, our interpretation is that verbal organization of the appropriate type affects the process of serial learning directly.

Two other problems worthy of further investigation are suggested by the present results. The first concerns the effect of sentential organization on the form of the serial position curve. That is to say, it is pertinent now to examine in more detail the process of facilitating response integration as reflected in the numbers of

items learned per trial and in the order in which they are learned. If the verbal context provided is critically involved in this process, variations in sentence properties such as phrase structure should affect the magnitude and location of errors in learning. Through the application of a phrase-structure analysis, Johnson (1965) has achieved a remarkable degree of success in predicting the error frequencies in the learning of sentences as responses in a PA task. A similar application might prove fruitful in the case of serial learning.

Finally, since it has been demonstrated that the provision of a verbal organization containing all of the items in a serial list facilitates learning, it is of interest to determine the conditions under which positive transfer would occur. One approach to this goal would involve the manipulation of both training and instructional variables relevant to the use and generation of verbal organization in the learning of serial lists. The effectiveness of the manipulations could then be evaluated in terms of performance on a transfer task administered in accord with the usual method of serial anticipation.

REFERENCES

JENSEN, A. R. Transfer between paired-associate and serial learning. *Journal of Verbal Learning and Verbal Behavior,* 1962, **1,** 269–280.

JENSEN, A. R. & ROHWER, W. D., JR. Verbal mediation in paired-associate and serial learning. *Journal of Verbal Learning and Verbal Behavior,* 1963, **1,** 346–352.

JENSEN, A. R., & ROHWER, W. D., JR. What is learned in serial learning? *Journal of Verbal Learning and Verbal Behavior,* 1965, **4,** 62–72. (a)

JENSEN, A. R., & ROHWER, W. D., JR. Syntactical mediation of serial and paired-associate learning as a function of age. *Child Development,* 1965, **36,** 601–608. (b)

JOHNSON, N. F. The psychological reality of phrase-structure rules. *Journal of Verbal Learning and Verbal Behavior,* 1965, **4,** 469–475.

(Received May 1, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 3, 191-196

MULTIPLE- VERSUS SINGLE-PROBLEM TRAINING AND VARIATIONS OF SOLUTION RULES IN THE FORMATION OF LEARNING SETS

FRANCIS J. DI VESTA AND RICHARD T. WALLS
Pennsylvania State University

The present study investigated the effects of multiple- vs. single-concept training orthogonally crossed with fixed, blocked, and scrambled solution rules on "learning to learn." Anagrams with scrambled solutions were more difficult to learn than those with single or blocked solution rules. Otherwise the manipulations of solution rules did not differentially affect later problem-solving proficiency. Multiple-concept training resulted in greater proficiency than single-concept training after the first anagram within a block was solved and facilitated transition to new sets of anagrams. However, single-concept training resulted in greater proficiency once the transition was made.

One facet of transfer that facilitates the rapid solution of problems is the acquisition of learning sets or "learning to learn" (LTL). Simply defined this means that when individuals learn they learn not only the specific concept, skill or discrimination demanded by the situation but they also learn something about *how* to form a concept, to develop a skill, to make precise discriminations, or to solve problems. A technical definition of LTL suggests that the improvement of performance is due to the similarity in the general processes fundamental to the learning of several successive tasks which are all representatives of a given class but which have no other systematic similarity among them (Underwood, 1966, p. 510). Harlow (1949), in his early study on learning sets with animal subjects (Ss), called attention to the importance of the principle of practice for this aspect of human learning. Through many practice trials on related problems the individual becomes proficient in solving new problems within a given class due to nonspecific transfer of learning to learn.

Because of its pervasive nature LTL is one of the important outcomes of the educational process. Unfortunately, it may also be one of the most neglected learning outcomes in compensatory education (Bloom, Davis, & Hess, 1965) as well as in most formal school settings despite the many innovations in content made through the so-called curriculum reform.

An analysis of the practice conditions for LTL suggests variables that ultimately may prove to be of importance for teaching methodology (see also Di Vesta & Walls, 1967a, 1967b, 1967c). As Adams (1954, p. 15) indicates, Harlow's method of training on a large number of problems may be only one of several possible training techniques that might be used. Thus, the successive tasks in initial training might be repeated presentations of the same problem, or they might be presentations of multiple problems. When working with concepts this implies a comparison of solution times for problems in which the same concept is employed with solution times for problems in which multiple concepts are employed. Furthermore, practice on problems blocked according to concept could lead to greater facility in the solution of new problems than practice with a variety of concepts varied unsystematically. Finally, the method of solution might be fixed for all problems; different for each set of problems but the same for trials within a problem; or different for all trials within a given problem.

The present study investigated the effects of certain of these variables on problem solving. In particular, this study evaluates the relative effectiveness of single-concept and multiple-concept training; and of fixed, blocked, and scrambled solution rules, during training, on time required for solution of successive sets of anagrams comprising the transfer task. The use of anagrams permitted adherence to the characteristics

of LTL, as described by Underwood (1966, p. 510) that (*a*) successive sets of problems represent samples of items of the same class of materials, and (*b*) no systematic similarity exists between the successive sets of problems other than the commonality which allows them to be said to be of the same class of materials.

METHOD

Design

One part of the design consisted of two variations in stimulus conditions operationally defined as the concept class. In the same-concept (SC) condition all anagrams in the training phase belonged to the same concept class. In the multiple-concept (MC) condition, *S*s were trained on several concept classes. These stimulus conditions were crossed orthogonally with three variations in response (responses were defined as solution rules) conditions. In the fixed-response (FR) treatment the solution was fixed, that is, the same solution rule could be used to unscramble any anagram in the entire training series. In the blocked-response (BR) condition the solutions were grouped so that the same solution rule could be used within a given problem but the solution rules differed among problems. The same solution rules used in the BR condition were also used in the mixed-response (MR) condition. However, in the MR condition the solution rules were assigned at random to the anagrams.

Within these primary cells of the design there were five sets (blocks) of anagram problems. Within each block there were seven anagrams (trials). This organization of five blocks of problems with seven trials in each block was employed in *both* the *training* and *transfer* series. Thus, each of the six groups of *S*s were required to solve a total of 70 anagrams.

Analyses were made of improvement in performance over blocks of trials and of improvement in performance as a consequence of position within blocks. For these purposes, the 2 × 3 × 5 × 7 design was analyzed by Lindquist's (1953) extended Type VI (two between and two within dimensions) mixed analysis of variance, separately for the training and for the transfer series.

Subjects

A total of 90 *S*s, 15 in each of the six conditions, participated in the study. The *S*s were undergraduate educational psychology students at Pennsylvania State University. There were 25 males and 65 females in the total group. The *S*s were assigned randomly to the six conditions with the restriction that these assignments be balanced over all cells. There were no restrictions placed on the selection of *S*s, many of whom had participated in previous verbal learning experiments. However, none had more than casual experience with anagram problems.

Materials

The materials for the SC conditions consisted of 35 five-letter anagrams of words from one concept class, that of articles of personal wear such as gloves, scarf, watch, and smock. The materials for the MC conditions were 35 five-letter anagrams of words from the five categories of musical instruments (viola, cello, flute, etc.), flowers (lilac, phlox, etc.), nationalities (Swede, Dutch, etc.), parts of the anatomy (ankle, chest, etc.), and whiteness (paper, chalk, etc.).

All anagrams for the FR conditions could be solved by employing the solution rule of 5,2,1,4,3. The solution rules for the BR and MR conditions were 1,3,5,2,4; 4,3,2,1,5; 1,3,5,2,4; 3,5,2,4,1; and 3,4,5,1,2. These sets of solution rules were applied at random to the blocks of words when constructing the anagrams. As noted earlier the solution rules for the BR condition were varied so that one rule could be used to solve all anagrams within a block but each block was solved by a different rule. The five solution rules for the MR conditions were assigned randomly throughout the series of anagrams.

All groups solved the same transfer task. This task was blocked into five categories of seven anagrams in each block, as noted above. The categories were animals (horse, mouse, etc.), fruit (pears, apple, etc.), furniture (table, stove, etc.), colors (black, white, green, etc.), and meat (bacon, steak, etc.). A different solution was used for each block of trials. Although none of the concept categories or solution rules were the same as those used in the learning task, the transfer task was essentially a continuation of the MC-BR (multiple-concept and blocked-response) condition in the initial learning task.

The length of all words was five letters and all words were nouns. The categories and total lists were also matched as closely as possible on digram frequency counts (Mayzner & Tresselt, 1965). In order to counteract the possibility that some of the anagrams were easier to solve than were others, three lists were formed for each condition by three random assignments of the selected words within each block. This process was followed for all of the learning and transfer task lists.

The anagrams were typed in capital letters in the center of white unlined 3 × 5 inch index cards. The cards were arranged in decks.

Procedure

The *S* was seated opposite the experimenter (*E*) at a small table. After brief introductory remarks *E* instructed *S* by instructions adapted from Ronning (1965) as follows:

> We are interested in finding the typical performance on several word problems. I want you to try to do your very best and to solve the

problems as rapidly as possible. We will be working with what is called the anagram problem, and with five-letter anagrams in particular. A five-letter anagram is simply five scrambled letters which when properly rearranged will form one five-letter word.

The S then solved five practice anagrams each with different solution rules. Any questions were answered by E. The solutions were timed to adapt S to the sound of the click of the stopwatch but the time was not recorded on the practice problems.

In order to compensate further for any differences in difficulty among different parts of a given list, each S was started at different points on both the training and transfer lists. For this purpose, one of the five blocks was selected by consulting a table of random digits. Each S then solved the 70 anagrams required for the treatment to which he had been assigned. He reported his answer (the noun) for each anagram and then slowly spelled it, touching each letter with the eraser of his pencil as he did so. The instructions made no reference to the solution rules. If S found an "incorrect" word solution, he was told to "find another word," and the time was reset to zero. A correction procedure was employed in those cases where S was unable to find the solution to any anagram within 60 seconds.

For experimental and analytical purposes each list in both the training and transfer series was viewed as consisting of five blocks of trials with seven anagrams (trials) in each block. It should be noted, however, that S was presented the materials for a given condition as a single continuous series of tasks proceeding, without a break or additional instructions, from the initial anagram in the training series to the final anagram in the transfer series.

RESULTS

Performance in the learning phase and in the transfer phase was measured by time to solution, in seconds, for each anagram. The data were analyzed, by a mixed analysis of variance with two between and two within dimensions, separately for the two phases.

The analysis of the data for the training phase yielded significant main effects due to variations in the solution rules ($F = 3.95$, $df = 2/84$, $p < .05$), to blocks of trials ($F = 6.67$, $df = 4/336$, $p < .01$), and to anagram position within blocks of trials ($F = 17.75$, $df = 6/504$, $p < .01$). In addition, the effect due to the interaction between anagram position and variation in the concept conditions was significant ($F = 14.76$, $df = 6/504$, $p < .01$). None of the other main effects or interactions was significant ($p > .05$).

The mean number of seconds to solve the anagrams by Ss in the BR, FR, and MR conditions were 23.13, 24.54, and 29.16 seconds, respectively. The average time required to solve anagrams in each of the seven positions by Ss in the MC condition compared to the average time required by Ss trained in the SC condition are displayed in Figure 1. (Note that the means in that display are based on data for a given position summed over all training- or transfer-phase trials.) Although there is an overall decrease in solution time (see solid line in Figure 1), it can be seen that single- and multiple-concept training apparently result in different strategies for the solution of anagrams. As might be expected, multiple-concept training results in longer time to solve the anagram in the first position. Once solved, and the concept is identified, the remainder of the anagrams in that concept category are solved quickly. The increase in time to solution from Position 5 to Position 7 by the MC group is undoubtedly due to the fact that anagrams are blocked according to concept; that is, Ss eventually learn that there are limited numbers of instances for each concept but have not in the course of the training phase discovered that it is seven. Presumably, in anticipation of a change in concept Ss in the MC group take slightly longer time to solve the last two anagrams in a series than for the anagram in the fifth position.

The analysis of variance for the transfer phase data yielded significant main effects due to anagram position within blocks ($F = 34.22$, $df = 6/504$, $p < .001$), to blocks of trials ($F = 3.51$, $df = 4/336$, $p < .01$), and a significant interaction between blocks of trials and kind (SC versus MC) of initial training ($F = 3.45$, $df = 4/336$, $p < .01$). None of the other main effects or interactions was significant ($p > .05$). The mean performance of all Ss on the anagrams in each position within the concept series in the transfer phase is displayed in Figure 1. The respective mean performances for the MC and SC training groups on this aspect of the transfer phase are only decimals apart

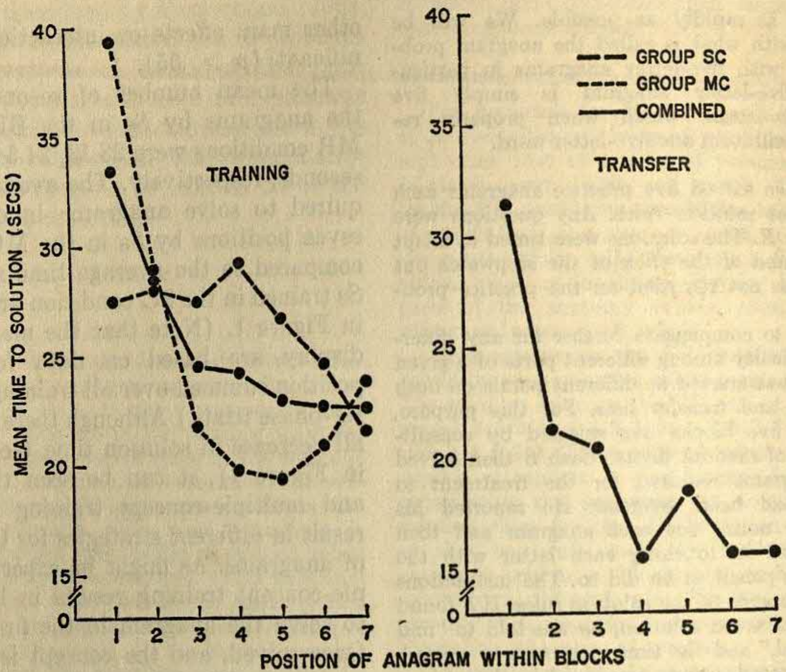


FIG. 1. Mean time to solve an anagram as a function of its position within blocks of seven trials during the training and transfer tasks.

and were not plotted. Consequently, it is interesting to note that the slight increase in time taken on the anagram in fifth position of the transfer series is characteristic of both groups. This effect appears to be similar to the increase found for the MC group on the training task although it is less pronounced in the transfer task.

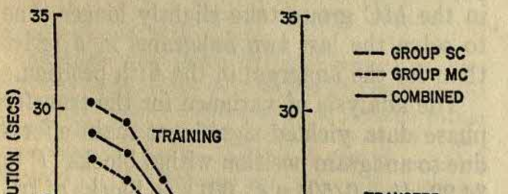


FIG. 2. Anagram solution times for five successive blocks of trials in the training and transfer tasks.

The data for the significant interaction between kind of concept training and blocks of trials are depicted, graphically, in Figure 2. As can be seen in that display there is a pronounced improvement in overall performance from the initial training trial to the final test block of trials. The differences between the SC and MC training groups were not significant ($p > .05$) on the training task but are presented here to permit comparisons by the reader. It can be seen that the performance of the MC training group is slightly superior over all but the fourth block of trials to that of the SC training group during the training task. This superiority is maintained on the first block of trials in the transfer task, as we might expect, since *S*s in the SC group have been trained on single concepts. On the remaining blocks of trials the performance of *S*s in the SC group is equal or superior to that of *S*s in the MC group.

The differences in the performance of the SC and MC groups on the transfer task suggest that quite different LTL processes

are being learned. While it would appear that MC training should lead to more positive transfer than SC training this does not appear to be true for the specific conditions of this experiment. It seems quite likely that in the present MC training conditions Ss learn situation-specific rather than task-related problem-solving strategies. Thus, for example, the MC training group might have learned strategies related to the arrangement of concepts in blocks of seven rather than of the more durable problem-solving strategies of attempting alternate moves, identifying syllables, searching for associates of bigrams and trigrams, and the like.

DISCUSSION

In this study, the effects of anagram training, with different conditions of concept categorizations and solution rules, on transfer to the solution of new anagrams were investigated. Manipulation of solution rules was found to affect performance during training only and did not differentially affect performance in the transfer task. The significant difference obtained during training simply reflects the greater difficulty in solving anagrams with inconsistent solution rules as opposed to consistent solution rules whether constant throughout all problems or whether blocked according to concept categories.

The principal effects on learning and transfer were due to manipulations of concept categories. MC and SC training had differential effects, during the training phase, on the solution of anagrams according to their positions within blocks of trials. Thus, during training the MC groups took longer to solve the first anagram within a block but solved the remainder more rapidly than did the SC group, thereby providing overall superiority to the MC group as measured by average time taken to solve all anagrams. However, since this effect is to be found in the transfer phase for all groups, without differentiation between groups, it must be concluded that the position effect is due to the manner in which the concepts were blocked. The adoption of different strategies for solving problems is an important outcome of the two training procedures. The MC training encourages the identification of the concept class on the first trial which then leads to the easy identification of associates of that class for the remaining anagrams within the class. SC training provides less practice on the "win-stay, lose-shift" strategy, learned in MC training, but provides more practice in identifying words belonging to a concept class.

The performance over all blocks of trials by the MC and SC training conditions compares favorably with classical demonstrations of learning to learn. However, the two types of training differentially affected transition to the transfer task. Thus, if Ss were trained on a single category of anagrams, a "set" was established that interfered with the immediate solution of a new class of problems. Nevertheless, after the first block of trials on the transfer task their performance was equal or superior to that of Ss trained on MC classes. As indicated by Johnson (1966),

> The development of a category set resembles the conventional concept learning experiment in that there are abstract similarities and superficial differences between successive problems and that, as practice continues, the similarities are more readily perceived and a common pattern of response follows [p. 373].

The present results are similar to those found by Adams (1954) who employed a much simpler training procedure than used in the present study. The advantage of MC training appears to be its effects on the ease of transition to a new problem. As with the comparable group in Adams' (1954) study, the performance of the SC group equaled or surpassed that of the MC group once the transition to multiple categories was made.

In conclusion, the trend of decreasing solution times over all blocks of trials suggests that increased proficiency in solving anagrams under the conditions of this investigation can be attained after experience with a sufficient number of training problems. Since all stimuli differ on bases other than concept classes, the effects are assumed to be due to the nonspecific transfer of learning how to learn.

REFERENCES

ADAMS, J. A. Multiple versus single problem training in human problem-solving. *Journal of Experimental Psychology*, 1954, **48**, 15–17.

BLOOM, B. S., DAVIS, A., & HESS, R. *Compensatory education for cultural deprivation*. New York: Holt, Rinehart & Winston, 1965.

DI VESTA, F. J., & WALLS, R. T. Response selection as a function of instructions and motivation under nonreinforcement conditions. *Journal of Experimental Psychology*, 1967, **73**, 365–373. (a)

DI VESTA, F. J., & WALLS, R. T. Transfer of object-function in problem-solving. *American Educational Research Journal*, 1967, **4**, 207–216. (b)

DI VESTA, F. J., & WALLS, R. T. Transfer of solution-rules in problem-solving. *Journal of Educational Psychology*, 1967, **58**, 319–326. (c)

HARLOW, H. F. The formation of learning sets. *Psychological Review*, 1949, **56**, 51–65.

JOHNSON, D. M. Solution of anagrams. *Psychological Bulletin*, 1966, **66**, 371–384.

LINDQUIST, E. F. *Design and analysis of experiments in psychology and education*. Boston: Houghton-Mifflin, 1953.

MAYZNER, M. S., & TRESSELT, M. E. Tables of single-letter and digram frequency counts for various word-length and letter-position combinations. *Psychonomic Monograph Supplements*, 1965, **1**, 13–32.

RONNING, R. R. Anagram solution times: A function of the "ruleout" factor. *Journal of Experimental Psychology*, 1965, **69**, 35–39.

UNDERWOOD, B. *Experimental psychology*. New York: Appleton-Century-Crofts, 1966.

(Received May 12, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 3, 197-201

SOME UNPREDICTED EFFECTS OF DIFFERENT QUESTIONS UPON LEARNING FROM CONNECTED DISCOURSE

LAWRENCE T. FRASE[1]
University of Massachusetts

It was predicted that a general orienting question would require that Ss process more information from a 36-word passage than would less general questions. Validity data on the question categories (60 college Ss) confirmed that the number of words in the passage which Ss thought were necessary to answer the questions increased from specific to general. In spite of pretraining, Ss given the general questions ignored words from the stimulus portion of the S-R pairs presented in the paragraph. Neglecting critical stimuli was considered to be an information rejection strategy which accounted for the results of another study, in which 84 Ss were instructed to use the questions as aids in learning the passage. Contrary to predictions, retention was lowest with general questions. Learning from connected discourse is interpreted as a multistage process which requires precise orienting instructions.

Anderson (1967) seems to have put his finger on a critical variable for controlling learning behaviors. He states that "... the most compelling stimulus in a frame is the question which must be answered or the blank which must be completed [p. 137]." A basic problem involved in the effective control of learning behaviors thus may not be whether the material is broken into small steps or physically separated stimulus and response terms, but whether the method of instructional control (be it a program frame, a question, or a graph) gets the student to practice the stimuli and responses, and to make the appropriate associations between the two. Learning from connected discourse, according to this model, would be similar to paired-associate learning in that more than one stage may be required to achieve mastery of the material (Underwood, Runquist, & Schulz, 1959). The point here is that there are several alternative ways of getting subjects (*Ss*) to go through the behaviors involved in these separate stages, and the use of questions is one of these ways.

The investigation of the instructional effects of questions is by no means a new experimental area (Distad, 1927; Holmes, 1931), yet very little precise information is available to tell us how questions work. In one study, Hershberger and Terry (1965) found that a confirmation procedure was least effective in a programmed learning task. The *Ss* who were given the correct answers (confirmation) presumably did not read the stimulus material carefully. These authors concluded that question difficulty (availability of the correct response) was an important determinant of learning. Rothkopf (1965) also found that if the correct response is easy to predict less will be learned. The most difficult questions evidently require *Ss* to process the words to which they are exposed since they learn more. The basic problem, however, is to determine what specific stimulus controls cause *Ss* to retain more when difficult questions are asked.

A study by Mechanic (1962), sheds some light on this problem. Using a paired-associate task, he found that the nature of the responses required by an orienting task (which related to different cues in the stimulus lists) was of crucial importance for learning. If the cues used in the orienting task (which might be questions) are relevant to the experimenter's (*E*'s) criterion, then *Ss* will score relatively high. Faust and Anderson (1967) found that making a program frame more difficult by adding irrelevant stimuli led to better retention because *Ss* had to at least notice the relevant stimuli. Another way of stating this would be that *Ss* were forced to

[1] Now at Bell Telephone Laboratories, Murray Hill, New Jersey.

respond discriminately to the stimuli when irrelevant components were added.

Questions can be used effectively because they function to control *S*s' attention. Berlyne (1965) maintains that attention is a negative process—it consists of information rejection. A precise question, such as asking for the name of an author, date of birth, etc., might allow *S* to ignore all but one sentence of a reading passage. If the question were more general, for instance, asking which of two authors was born earlier, *S* would have to take more of the information into account in order to answer the question. Schroder, Driver, and Streufert (1967) also emphasize the view that information rejection is a concomitant of attention. They point out that *S*s "filter" inputs (by rejecting certain information), and that more filtering will occur as the information load is increased. Information load may increase to a point at which *S*s will abbreviate the task at hand, adopting what may seem to be an optimal strategy. Under high load, for instance in a lengthy connected discourse task in which reading behaviors are not precisely controlled, overall retention of the passage would decrease if *S*s adopted a strategy which omitted some necessary step, such as practicing stimuli, practicing responses, or associating the two.

The problem explored in the present study was to determine what happens to retention of a passage when an orienting question is asked which requires processing a relatively large or small amount of the total information in that passage. Study 1 was conducted to determine, in a simple way, the validity of the question categories—whether *S*s agreed with *E* that more words are necessary to answer a general question as opposed to a specific question.

The basic hypothesis of this experiment, explored in Study 2, was that a general question, with a large number of associates within the reading passage, would require processing more information than a precise question, hence retention would be higher when *S*s received a general question before reading the passage. This hypothesis is consistent with the view that attention, under the control of certain questions, consists of rejecting irrelevant information.

To anticipate the results, the data of Study 2 were statistically significant—in the direction opposite to that predicted. The initial validity data from Study 1 showed that there was a plausible and interesting reason why this should have occurred.

STUDY 1

Method

Subjects. Sixty undergraduate educational psychology students participated in this study as a laboratory exercise.[2] The *S*s were randomly assigned to the three experimental groups.

Materials. A very simple, highly structured, paragraph of 36 words was constructed which described two attributes about each of four individuals. The paragraph follows.

> Jim is a pilot. He was born in 1921. John is a policeman. He was born in 1930. Jack is a butcher. He was born in 1926. Jeff is an engineer. He was born in 1934.

In addition, three sets of two questions were constructed which *E* labeled specific (S), comparative (C), and general (G). The specific questions were:

S1. When was Jack born?
S2. What does Jack do?

The comparative questions were:

C1. Is Jim older than Jack?
C2. Who has the more highly skilled job, Jeff or Jack?

The general questions were:

G1. When were the men in the paragraph born?
G2. What jobs do the men in the paragraph hold?

Hence, there was a specific, comparative, and general question related to age and occupation.

Procedure and design. In order to determine the validity of the question categories, *S*s were asked to underline all the words in the paragraph which would comprise *a complete* sentence giving the information needed to answer one of the six questions. An instruction sheet was constructed which described this task to *S* and which gave him an example and three practice problems (with knowledge of results) including all three types

---

[2] The author expresses special appreciation to Harry Schumer for providing the subjects.

of questions. The *Ss* were told that there was no time limit on this task. After reading the instructions *S* turned the page and found one of the six questions, below which was the paragraph. He then proceeded to underline the words in the paragraph.

The *Ss* from two laboratory sections were randomly assigned to one of six groups, 10 *Ss* received Question 1, 10 *Ss* received Question 2, etc., for a total of 60 *Ss*. The hypothesis was that *Ss* would underline most words with the general question, fewer with the comparative question, and least with the specific question. A simple one-way analysis of variance was planned.

Results

Table 1 presents the data which corroborate *E*'s characterization of the questions as specific, comparative, and general. The comparative and general questions were, according to *Ss*, associated with more words than the specific question. The increase in words from specific to general questions was significant whether or not connectives were included in the word count. There was no variance in the number of words underlined for specific and comparative questions, but there was for the general questions, hence the Kruskal-Wallis analysis of variance was used. There was very little difference between the two questions within each question category and the two questions (S1 and S2; C1 and C2; G1 and G2) were combined in this and later analyses.

The variability of *Ss*' scores with the general question was a troublesome finding. Obviously, the task was extremely simple, but for some reason *Ss* in the general question group seemed to ignore the instructions and practice problems they had been given. Assuming that *Ss* did what they were instructed to, they should have underlined one sentence if they had received a specific question, two sentences if they had a comparative question, and four sentences if they had a general question. To explore this idea further, the number of words which were not underlined when they should have been (extrusions) was tabulated for each question category. The data on extrusions in Table 1 suggest that the general question group for some reason adopted the strategy of throwing out words—presumably a form of information rejection.

TABLE 1

ANALYSIS OF WORDS UNDERLINED WHICH WOULD COMPRISE AN ANSWER TO THE QUESTIONS

Words	Specific		Comparative		General		H^a
	Mdn	Range	*Mdn*	Range	*Mdn*	Range	
Underlined	4.5	1.0	8.9	2.0	18.0	16.0	30.8**
Extrusions	—	—	—	—	1.4	17.0	7.7*

Note.—$N = 20$.
[a] Kruskal-Wallis analysis of variance (corrected for ties) was used.
* $p < .05$.
** $p < .001$.

It seemed fruitful to ask whether or not *Ss* were rejecting terms in any systematic manner. The data revealed two things. First, extrusions were confined entirely to the general question group, and second, within that group the extrusions were confined entirely to stimulus terms and connectives. Response terms (always the predicate of the sentence) are conceived as those terms which directly answer the questions. About 27% of *Ss* in the general question groups, instead of underlining the four sentences which were required simply underlined, "pilot," "policeman," "butcher," and "engineer," or the appropriate date. If *Ss* adopted this strategy when required to learn the paragraph it seemed doubtful that they would make the appropriate stimulus-response associations. It was expected that *Ss* might disregard connectives, but not the names of the men in the paragraph.

STUDY 2

Method

Subjects. Eighty-four *Ss* from three laboratory sections of educational psychology participated as a laboratory exercise. The *Ss* were different from those who participated in Study 1.

Materials and procedure. The same questions and 36-word paragraph used in Study 1 were used in this study. The following instructions were handed out to *Ss*.

> We would like you to read a paragraph which will be *followed by a test*. You will be provided with a test question which should aid you in getting relevant information from the paragraph.
>
> You will be allowed 20 seconds to read the paragraph.

When the experimenter says "Ready-turn!" turn over the page. You will see the question and paragraph. The experimenter will then say "Ready-turn!" again. At that time turn over the paragraph and DO NOT LOOK BACK.

The question, which was followed by the paragraph directly below it, was on a separate sheet of paper following the instruction page. After *S*s completed the reading task all papers were collected and the retention test was distributed.

A five-alternative multiple-choice test of nine items was administered which provided, as question stems, the names of the men in the paragraph. The *S*s had to select the correct date of birth and occupation for each name. There was one additional item which required *S*s to rank order men in terms of date of birth. In short, there was a test item for every sentence in the paragraph.

Design. The *S*s were randomly assigned to one of the three question treatments. There were 28 *S*s in each of the treatments (specific, comparative, or general question). Half of the *S*s in each of these three groups received the age-related question, the other half received the occupation-related question.

One hypothesis was that more *S*s in the specific question group would pass the one age or occupation test item which was relevant to their question. A chi-square test with a pass-fail criterion was planned to test this hypothesis concerning specific retention.

Another hypothesis was that the general question group would score highest over the entire test. A one-way analysis of variance was planned to test this hypothesis concerning generality of retention.

Results

The χ^2 test was significant at the .01 level ($\chi^2 = 10.8$, $df = 2$). Of the 28 *S*s in each group the percentage passing the one age or occupation item appropriate to the specific question group was 82% (specific), 61% (comparative), and 39% (general). The most precise question led to the most efficient acquisition of the specific stimulus-response association, confirming the hypothesis.

The other hypothesis to be investigated was whether general questions would lead to higher overall scores—generality of learning. The means for the specific, comparative, and general groups were 4.75, 4.25, and 3.25, respectively. For these data, $F = 4.35$, $df = 2/81$, $p < .05$. Duncan's multiple-range test indicated that only the specific and general group means differ at the .05 level. In short, the general question group scored lowest whether the criterion of performance was a specific question-relevant test item or the total retention test.

DISCUSSION

The results of both studies taken together suggest that questions can have subtle effects upon *S*s performance which may not be anticipated by the unwary instructor. Questions may work, in the sense that they cause *S*s to pay close attention to the passage, but the phrasing of the questions might select out only a portion of the necessary stimuli. In the present study the questions were designed to get *S*s to respond to a relatively small or large amount of the material, but the general questions did not work that way. Instead, several *S*s concentrated upon only the response terms. This finding may have implications for the inadequacy of *S*s' underlining or note-taking skills. On the assumption that learning from connected discourse involves both a response learning and associative phase (Underwood, Runquist, & Schulz, 1959), it is clear that questions which are to be used as instructional aids must be phrased in such a way that *S*s are directed to rehearse the stimuli, the responses, and also the associations between the two. In confirmation of the Faust and Anderson (1967) finding, an efficient frame (or question) must insure that *S*s practice more than just the response terms. To reword Anderson's (1967) maxim, the most compelling stimulus in a paragraph is the word or set of words which *directly* answers the question or fills in the blank. The precise responses required by the orienting task (questions), in confirmation of Mechanic's (1962) findings, seem to have been critical for learning in this study. The *S*s responses were more precise than *E*'s skill at making up prequestions.

The data on extrusions confirm the process of selective information rejection (attention) which Berlyne (1965) and Schroder, Driver, and Streufert (1967) have suggested. The general questions used in the present studies evidently represent a case of maximal information load or uncertainty. All *S*s received the same

paragraph, hence the amount of nominal uncertainty or information in the paragraphs was held constant, but the questions functioned to systematically induce more or less effective uncertainty into the learning task. With higher uncertainty (the comparative and general questions) the subtle inadequacy of the questions (in terms of the criterion retention test) became more critical, and *S*s adopted their own strategies of learning. The general conclusion seems to be that, as effective uncertainty or information load increases, precise control over reading behavior becomes more imperative. In a program frame information load is limited and hence precision of control is obtained by the format of presentation.

At the other extreme, learning from continuous discourse material (a high uncertainty condition) shifts the burden of control from the format of presentation to the orienting task. The orienting task, whether a question, a graph, or combinations of various aids, must insure that *S*s execute all the responses necessary for successful performance of the criterion task. This includes rehearsing the stimulus, rehearsing the response, and putting the two together. Ultimately such precise control reduces to a programmed learning task, except that it retains the advantage of keeping the learning material together in one place. Presumably, there are advantages in contiguous presentation of a topic (Ausubel, 1963). As this simple experiment has shown, problems of effective stimulus control become critical under free response learning conditions.

REFERENCES

ANDERSON, R. C. Educational psychology. *Annual Review of Psychology,* 1967, **18,** 129–164.

AUSUBEL, D. P. *The psychology of meaningful verbal learning.* New York: Grune & Stratton, 1963.

BERLYNE, D. E. *Structure and direction in thinking.* New York: Wiley, 1965.

DISTAD, H. W. A study of the reading performance of pupils under different conditions on different types of materials. *Journal of Educational Psychology,* 1927, **18,** 247–258.

FAUST, G. W., & ANDERSON, R. C. Effects of incidental material in a programmed Russian vocabulary lesson. *Journal of Educational Psychology,* 1967, **58,** 3–10.

HERSHBERGER, W. A., & TERRY, D. F. Delay of self-testing in three types of programmed text. *Journal of Educational Psychology,* 1965, **56,** 22–30.

HOLMES, E. Reading guided by questions versus careful reading and re-reading. *The School Review,* 1931, **39,** 361–370.

MECHANIC, A. Effects of orienting task, practice, and incentive on simultaneous incidental and intentional learning. *Journal of Experimental Psychology,* 1962, **64,** 393–399.

ROTHKOPF, E. Z. Some theoretical and experimental approaches to problems in written instruction. In J. D. Krumboltz (Ed.), *Learning and the educational process.* Chicago: Rand McNally, 1965.

SCHRODER, H. M., DRIVER, M. J., & STREUFERT, S. *Human information processing.* New York: Holt, 1967.

UNDERWOOD, B. J., RUNQUIST, W. N., & SCHULZ, R. W. Response learning in paired-associate lists as a function of intralist similarity. *Journal of Experimental Psychology,* 1959, **58,** 70–78.

(Received May 12, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 3, 202–206

EFFECTS OF WORDS AND PICTURES AS STIMULI IN LEARNING LANGUAGE EQUIVALENTS

STANLEY L. DENO
University of Delaware[1]

The present research was designed to contrast the psychological processes underlying perception of pictorial and verbal stimuli. 72 students learned a 12-pair set of either English word–Japanese word or simple picture–Japanese word equivalents which were either conceptually dissimilar, similar-isolated, or similar-grouped. The findings were that pictures and words representing the same familiar objects did not function as equivalent stimuli in learning the set of language equivalents. Generally, pictures facilitated the learning of equivalent pairs, especially when the objects represented by the abstractions were conceptually similar ($p < .005$). Further, grouping the abstractions representing conceptually similar objects increased the rate of acquisition.

Whenever the conditions for learning a semantic relation are arranged, a decision is made as to how the referent shall be presented. Although the choice may be made between concrete and abstract forms (i.e., between the events and their abstractions), within educational settings the choice more frequently is made between types of abstraction (e.g., between words and pictures).

In foreign language learning it has long been considered expedient, if not optimum, to learn the meaning of words through explicitly relating new words to old. Thus, vocabulary learning in a foreign language has often occurred by pairing a word or phrase from the to-be-learned language with its equivalent within the natural language under conditions where the learner understands that the foreign word "means the same" as its natural language equivalent.

Current trends in the teaching of foreign languages do not favor the word-word equivalents approach to learning meaning, however. Advocates of the "audio-lingual," "audio-visual," or "conversational" approach to foreign language learning (Rivers, 1964) contend that the learning of equivalents "binds" the learner to his natural language so that utterances within the new language must be preceded by the equivalent within the natural language. Such a two-step process, it is claimed, necessarily impedes the development of facility within the new language. Recent theoretical formulations (Deese, 1964, for example) and empirical evidence (Jenkins, Neale, & Deno, 1967; Karwoski, Gramlich, & Arnott, 1944) suggest, however, that nonlinguistic events are processed linguistically. If this is true, it may be virtually impossible to avoid the "two-step process" with a native speaker.

The present research was designed to contrast the psychological processes underlying perception of pictorial and verbal stimuli. Underwood and his associates (Underwood & Schultz, 1961; Wallace & Underwood, 1964) have clearly demonstrated that the rate of learning a set of word pairs is decreased when the members of the set are conceptually similar. The strategy employed in the present study was to use sets of words and pictures as stimuli which represented the same common objects under conditions of both high and low conceptual similarity. The prediction was that an increased difficulty in learning with a high similarity set of words would be paralleled by an increased difficulty in learning with a high similarity set of pictures. This hypothesis was based on the assumption that during learning

[1] The data on which this paper is based were included in the author's dissertation, presented in partial fulfillment of the requirements for the PhD degree at the University of Minnesota. Gratitude is expressed to Daniel C. Neale who directed this study.

the pictures would be linguistically encoded and processed in the same manner as the words.

METHOD

Subjects

Thirty-six males and 36 females enrolled in introductory educational psychology classes at the University of Minnesota during the summer of 1965 served as subjects (*S*s). The majority of the students were juniors enrolled in the College of Education. Each student participated in the experiment on a voluntary basis, but received course credit for his participation as an *S*.

Two restrictions were placed on the selection of *S*s. First, no student could serve as an experimental *S* if his native tongue was not the English language. Second, no student could serve as an *S* who had previously studied an Oriental language. The first restriction was applied to ensure that all *S*s would possess approximately the same natural language habits in relation to the stimuli employed, and the second restriction was applied to avoid the possibility that an *S* might have been familiar with the Japanese words which were learned as responses.

Materials

The *S*s learned a set of equivalents consisting of 12 pairs. The stimuli were either pictures or words representing 12 common objects and the responses were 12 Japanese words.

Twenty-four different objects were represented either by verbal (one word) labels or by simple black and white lined drawings. Twelve of these representations comprised a conceptually dissimilar list, and 12 comprised a conceptually similar list. The conceptually similar lists contained three instances from each of four conceptual categories—animal, clothing, furniture, and building. The nature of the lists employed can best be understood by examining Table 1.

As shown in the table, the conceptually similar list was arranged in two different ways. Either the instances from a particular category were presented in such a way as to be maximally separated from other instances within the same category during the presentation (List 2), or the category instances always appeared in sequence during the presentation (List 3).

The Japanese words used as responses were taken from the list of responses used by Horowitz and Larsen (1963). Japanese words were used as responses because their English transliterations are easily read and pronounced by persons native in the English language. Despite this relative ease of pronounciation, however, Japanese words are not derived from a language related to English, and, consequently, appear very unfamiliar to someone who has not had previous experience with them.

TABLE 1
REFERENT STIMULI AND JAPANESE WORD RESPONSES

Stimuli (either word or picture)			Responses
Dissimilar	Similar-isolated	Similar-grouped	
WOMAN	HAT	HOUSE	Atsui
DOOR	CHAIR	SCHOOL	Hune
FISH	CAT	CHURCH	Riko
KNIFE	TABLE	BED	Kari
NOSE	CHURCH	CHAIR	Baka
STOOL	DOG	TABLE	Amai
BOAT	SCHOOL	COAT	Hikui
MOON	TIE	HAT	Hayai
STORE	MOUSE	TIE	Kuro
APPLE	COAT	DOG	Tako
RADIO	HOUSE	MOUSE	Chikai
LADDER	BED	CAT	Tooi

The same 12 Japanese words were used as responses regardless of the objects represented by the stimuli. Using the same response words allowed comparisons among stimulus lists which were not confounded by different rates of response acquisition. The list of responses also appears in Table 1.

Two different random pairings of stimuli and responses were used to reduce the possibility that a particular set of stimulus and response pairings was easier to learn. The only constraint placed upon the randomizations was that a response was not paired with the same stimulus in both randomizations. The responses were, of course, paired with different stimuli in the cases of dissimilar and similar lists, but the responses were paired with the same stimuli for the two different arrangements of similar stimuli.

Three different random orders of the lists were made to reduce the possibility of sequence effects in response learning. The randomizations of Lists 1 and 2 were freely accomplished, but the related stimuli in List 3 had to be grouped so that randomizations for this list were composed by first randomly assigning the category positions within the list, and then randomly assigning the instances within each category.

Apparatus

All of the stimuli and responses were photographed on 35-mm. black and white film and the negatives were slide-mounted. The images thus produced were white on a black background. The white on black contrast was selected to reduce glare produced by rear projection. Two juxtapositioned images were projected at approximately eye level on a 15-inch × 30-inch rear projection screen by two carousel-type slide projectors.

The slide projectors were connected to a timer which advanced the carousel and presented a new slide every 4 seconds. On-screen images were

controlled by a shutter in front of each projector lens. The stimulus slides were placed in one projector and the response slides in another. As soon as the slides had advanced, a shutter in front of the stimulus slide would raise to present the stimulus (word or picture). After 2 seconds the response slide (Japanese word) would be exposed. Then, both shutters would drop and the timer would advance the carousel so that a new stimulus and response pair were available for presentation. During the advance the screens were black for slightly less than 1 second.

Procedure

The *Ss* participated individually. Each *S* was randomly assigned to 1 of 6 treatments based on the order in which he reported to the laboratory. The only restraint on this random assignment was that an equal number of male and female *Ss* should participate in each of the treatments. The treatment groups comprised a 2 × 3 factorial with two types of stimuli and three arrangements of list similarity as the conditions.

Each *S* was seated at a table before the rear projection screen and given instructions that he was going to learn the meanings of some Japanese words. It was explained to him that he was to learn these meanings under a paired-associate anticipation method. First the event signified by the word would be projected on the screen and then a few moments later the Japanese word which was equivalent to that event would be flashed alongside of it. His task was to learn the meaning of each word so that on succeeding presentations of the abstraction he would be able to anicipate the Japanese word before it flashed upon the screen. The *Ss* were encouraged to guess if they were not sure.

The three random orders of each list were then presented continuously to *Ss*, although there was a brief delay following Order 3 so that the carousel could be advanced to the beginning of Order 1 again. The *Ss* continued through the list until they correctly anticipated the Japanese word responses on two consecutive trials, or until 26 trials had occurred.

Recording and Treatment of Data

All responses were recorded by the experimenter on a data collection form. Comparisons between groups were made in terms of number of trials to criterion (twice through the list without error) and error rates. Since different stimulus lists might require more trials for learning, and, therefore, increase the number of opportunities to make errors, total errors was not used as a basis for comparison. Instead, the number of errors in 10 trials was used to compare error rates. The first 10 trials were selected because beyond that point many *Ss* attained criterion.

Errors were classified as omissions (failures to respond), and intrusions (overt errors), and separate comparisons were made for both types as well as omissions and intrusions combined.

RESULTS

The results of learning the equivalent word-word and picture-word pairs are summarized in Table 2. The error means in the table are based upon responding during the first 10 trials. Six *Ss* in the similar-isolated words group failed to achieve the criterion within 26 trials and were given a trials-to-criterion score of 28 (the fewest number of trials in which these *Ss* could have achieved the criterion).

The results are reported in terms of a set of orthogonal comparisons which were of prior interest.

TABLE 2

MEANS, STANDARD DEVIATIONS, AND PROBABILITY VALUES FOR DIFFERENCES ON TRIALS TO CRITERION[a] AND ERROR MEASURES[b]

Stimulus Condition	Trials to criterion			Omission errors			Intrusion errors			Combined errors		
	$\bar{X}$	*SD*	*p*	$\bar{X}$	*SD*	*p*	$\bar{X}$	*SD*	*p*	$\bar{X}$	*SD*	*p*
Dissimilar												
Words	15.2	6.4	—	43.7	20.8	—	14.5	10.0	—	58.2	13.1	<.25
Pictures	14.2	5.0		38.3	17.7		9.8	9.5		48.2	20.2	
Similar-isolated												
Words	23.8	5.2	<.0005	71.0	13.0	<.005	17.3	9.8	—	88.0	13.5	<.01
Pictures	15.9	4.8		47.8	20.5		19.7	17.4		67.5	18.9	
Similar-grouped												
Words	17.7	6.1	<.025	50.8	15.9	—	13.7	11.0	—	64.5	18.8	—
Pictures	13.4	4.7		44.4	13.0		13.2	10.0		57.6	14.8	

[a] Twice through the list without errors.
[b] Errors based on responding in the first 10 trials.

Dissimilar Stimuli

The results of learning the equivalent pairs where the stimuli were conceptually dissimilar show that *S*s learning with pictures performed better than *S*s learning with words on every dependent variable. In no case, however, are the obtained differences between group means reliable.

Similar-Isolated Stimuli

Isolating similar stimuli in the sequence of presentation produced very large and significant advantages for learning with pictures rather than words as stimuli on all dependent variables except intrusion errors, ($F < 1.00$, $df = 1/66$). Whereas the mean number of trials to criterion increased from 14.2 with dissimilar pictures to only 15.9 with similar-isolated pictures, the mean number of trials with similar-isolated words as stimuli increased from 15.2 to 23.8. The large difference in error rate between pictures and words which were similar and isolated was due to the difference in errors of omission, ($F = 11.01$, $df = 1/66$, $p < .005$). Within the first 10 trials *S*s in the similar-isolated pictures condition actually made more intrusion errors than *S*s in the similar-isolated words condition (words $\bar{X} = 17.3$: pictures $\bar{X} = 19.6$), although the difference was very small and not statistically significant ($F < 1.00$, $df = 1/66$). In sum, *S*s learning with words or pictures were about equally likely to make an overt response which was *incorrect*, but *S*s learning with pictures were more likely to make overt responses which were *correct* than *S*s learning with words. The overall effect was that those *S*s learning with similar-isolated pictures emitted Japanese word responses at a much higher rate during the first 10 trials than those *S*'s learning with similar-isolated words. This, in spite of the fact that all *S*s had an equal opportunity to acquire and emit the responses during learning.

Similar-Grouped Stimuli

Grouping similar stimuli during the sequence of presentation produced differences between word and picture means which were much smaller than with similar-isolated stimuli. As measured by trials to criterion the picture-word difference for similar-grouped stimuli, although statistically significant ($F = 5.88$, $df = 1/66$, $p < .025$) was approximately one-half the difference obtained between pictures and words when similar stimuli were isolated during presentation. The difference between words and pictures which were similar but grouped during presentation was statistically significant only for number of trials to criterion. None of the error rate measures yielded a mean difference which approached statistical significance ($F < 1.00$, $df = 1/66$ in all cases).

DISCUSSION

The results indicate that it makes little difference whether words or pictures are used to represent objects when learning a set of equivalent pairs. The conclusion holds, however, only if the objects represented are conceptually dissimilar. When the events represented are conceptually related, pictures seem to be much more easily associated with foreign word responses than words. The extent of advantage of pictures over words is different, however, depending upon the dependent variable considered, and whether or not conceptually related stimuli are presented in isolation or in groups during learning. In terms of trials to criterion, learning with conceptually related words is significantly more difficult than learning with conceptually related pictures regardless of stimulus arrangement. With respect to error measures, however, statistical significance depends upon the sequence in which conceptually related stimuli are presented to the learner (i.e., whether isolated or grouped).

The findings with respect to grouping conceptually similar stimuli are consistent with the results of a study by Gagné (1950) in which grouping similar nonsense form stimuli facilitated learning. In the present study, grouping conceptually related stimuli reduced learning difficulty to a point where a set of pairs could be acquired al-

most as well with similar stimuli as with dissimilar stimuli.

The obtained interaction between stimulus mode and stimulus similarity is particularly interesting. If it is assumed that the increased difficulty in learning with conceptually related words occurs because of similarity in meaning (conceptual similarity), then it must be concluded that the pictures, although readily identified with the appropriate word labels, are not encoded in the same manner as the words. This conclusion derived from the differential effects obtained during the learning of language equivalences (i.e., in a paired-associate learning task) is consistent with evidence obtained by Deno, Johnson, and Jenkins (in press) in a study comparing the distributions of free associations to words and pictures. These investigators, using the same words and pictures as in the present study, found that mode of representation (word or picture) significantly altered the associative similarity between objects represented.

Both the differential efficiency in learning with words and pictures and the apparent difference in psychological effect are seen as relevant for any one attempting to use these two different kinds of abstractions to communicate meaning. These results are particularly significant in foreign language learning. While learning the referent for a foreign word may be more efficient if the event is portrayed pictorially, there is the danger that the meaning evoked by the picture may not be the same as that intended. It may be that communication by word is more reliable.

REFERENCES

DEESE, J. Effects of instructions to learn. In A. Melton (Ed.), *categories of human learning.* New York: Academic Press, 1964, pp. 202-209.

DENO, S. L., JOHNSON, P., & JENKINS, J. R. The structure of associations to words and pictures. *A-V Communications Review,* in press.

GAGNÉ, R. M. The effect of sequence of presentation of similar items on the learning of paired associates. *Journal of Experimental Psychology,* 1950, **40,** 61-73.

HOROWITZ, L. M., & LARSEN, S. R. Response interference in paired-associate learning. *Journal of Experimental Psychology,* 1963, **65,** 225-231.

JENKINS, J. R., NEALE, D. C., & DENO, S. L. Differential memory for picture and word stimuli. *Journal of Educational Psychology,* 1967, **58,** 303-307.

KARWOSKI, T. F., GRAMLICH, F. W., & ARNOTT, P. Psychological studies in semantics: I. Free association reactions to words, drawings, & objects. *Journal Social Psychology,* 1944, **20,** 233-247.

RIVERS, W. A. *The psychologist and the foreign language teacher.* Chicago: University of Chicago Press, 1964.

UNDERWOOD, B. J., & SCHULZ, R. W. Studies of distributed practice: XXI. Effect of interference from language habits. *Journal of Experimental Psychology,* 1961, **62,** 571-575.

WALLACE, W. P., & UNDERWOOD, B. J. Implicit responses and the role of intralist similarity in verbal learning by normal and retarded subjects. *Journal of Educational Psychology,* 1964, **55,** 362-370.

(Received June 15, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 3, 207-214

PART-TASK VERSUS WHOLE-TASK PROCEDURES FOR TEACHING A PROBLEM-SOLVING SKILL TO FIRST GRADERS[1]

RICHARD C. ANDERSON
University of Illinois

Groups of 1st graders were trained to solve concept-attainment problems by either a small-step, programmed part-task method or a whole-task method in which S attempted terminal problems early in training. Contrary to previous research, which had suggested that whole-task methods are superior for highly organized tasks, in the present instance the part-task group performed better than the whole-task group on terminal training problems and on similar problems presented again later to measure retention; however, there was no difference between these groups on transfer problems. Both training groups were superior to a no-treatment control group on all measures.

Educators who have been influenced by the programed-instruction movement take it as self-evident that the best way to teach a complex skill is to analyze it into component subskills and subconcepts, then teach each of these in turn. Cast in different language such an approach is a part-task method, to be contrasted with the whole-task method in which the student is required to perform the terminal behavior as best he can from the very beginning of training. Surprising as it may seem to those who have been influenced by the conceptions of programmed instruction, the research on complex skill training has frequently shown whole methods to be superior to part methods.

The terms "part" and "whole" will be used in this paper as shorthand words for talking about the issue of how lengthy and complicated a segment of a task the student should be required to attempt during instruction, especially during the initial stages of instruction. Part methods result in low initial error rates and fast progress, at least at the beginning of instruction, but when account is taken of the time to combine the parts, typically the advantage for the part method has been negligible at best. In the case of rote materials, practice on later parts (sublists) produces interference with earlier parts. This interference must be overcome during the combination stage. With respect to complex skills and structured, meaningful material, there are coordinations and interrelationships among the subskills and subconcepts that cannot be acquired from training with the components alone. Herein lies one apparent reason that whole-task training has frequently proved superior to part training in the case of complex skills.

Whether a procedure which emphasizes lengthy task segments will prove superior to a procedure that begins with short task segments will surely be heavily dependent upon the manner in which the training procedures are developed. If entering behavior is underestimated and the steps in the part procedure are more finely granulated and more numerous than necessary, it may be less efficient than a procedure in which larger task segments are emphasized. Further, particularly when the task analysis underlying the part procedure is incomplete, subjects (Ss) who receive the part or small-step procedure may fail to learn coordinations among component skills and concepts, whereas Ss trained from the beginning on larger segments of the task may induce these coordinations.

There may also be characteristics of tasks which systematically interact with

[1] The research described herein was completed under a grant from the Office of Education, United States Department of Health, Education, and Welfare. The author is indebted to Donald Cunningham, Diane Fuglsang, Marianne Roderick, and Rivka Yavin for their assistance in this project. The author is grateful for the cooperation of the staff of South Side School, Champaign, Illinois; Tolono School, Tolono, Illinois; and Prairie School, Urbana, Illinois.

the length and complexity of the instructional units into which the task is divided. Naylor and Briggs (1963) have suggested that the relative effectiveness of part and whole methods is a function of the level of complexity and the degree of organization of the task. Task complexity is said to refer to "demands on information-processing and memory-storage capacities" while task organization is said to refer to the nature and extent of the interrelationships among task dimensions. As organization increases, whole methods are predicted to be increasingly superior to part methods. For a highly organized task, an increase in complexity (difficulty) is predicted to result in greater superiority for the whole method. The part method is predicted to be superior to the whole method only in the case in which the task is both complex and unorganized. Naylor and Briggs (1963) completed an experiment, using what may be called concept-learning tasks, in which it was found that the whole method was much better than the progressive-part method on a highly organized task regardless of task complexity. On an unorganized task, the whole method was slightly better when task complexity was low and very slightly worse than the progressive-part method when task complexity was high.

The experiment reported herein involved a comparison of a small-step, programed part-task method and a whole-task method for teaching children a complex problem solving skill, the skill of varying each factor in succession while holding all other factors constant. This skill is a classical strategy of experimental science and it is applicable to a large and important class of problems (Bruner, Goodnow, & Austin, 1956, pp. 81–125).

A concrete example will serve to illustrate the kind of problem presented to the child and the sort of behavior expected from him. The materials consist of eight cards upon which are pictured either one or two rectangles or diamonds which are either red or green. The following instructions are read to the child:

> I will think of a secret and point to one card that shows my secret. You pick cards to figure out my secret. Each time you pick a card I will tell you whether it shows the secret. As soon as you know, tell me the secret.

There are 27 different conjunctive concepts which can be formed with the card array (and the other materials described later). There is 1 zero-dimensional concept (All the cards) and there are 6 one-dimensional concepts (for example, red or two), 12 two-dimensional concepts (for example, green diamond), and 8 three-dimensional concepts (for example, one red rectangle). The problem might involve any one of the 27 possible concepts. The child does not know which stimuli are discriminative nor how many dimensions are relevant. He begins with only a positive or focus instance. It is not enough that the child says the right answer. To be counted as having solved the problem, the child must choose a set of cards and state a concept such that the set of cards logically implies the concept that he states and no other.

Relative to the tasks employed by Naylor and Briggs (1963) the problem-solving task used in the present experiment would have to be regarded as highly organized. Furthermore, the task is very difficult (complex) for seven-year-olds (Anderson, 1965), the age of the *S*s in the present study. According to Inhelder and Piaget (1958, p. 335) people do not normally acquire the skill taught in the present experiment until 14–15 years of age. For these reasons, if Naylor and Briggs (1963) are correct, this should be a case in which the whole method is vastly better than the part method.

METHOD

Materials

Three sets of materials were employed. The card array consisted of eight 2½ × 3¼ inch cards taped in an orderly arrangement on a 16 × 18 inch Masonite panel. The cards had figures inscribed upon them that varied with respect to number (one figure or two figures), color (red or green), and form (rectangle or diamond). The cowboy game consisted of eight toy plastic cowboys and accessories imbedded in an orderly arrangement

in a 12 × 13 inch plaster-of-paris base. The cowboys were either standing or riding horses, either with or without hats, and either with or without rifles. The final set of materials consisted of eight yellow, hexagonal No. 2 pencils. The pencils displayed three attributes: length (6 or 2 inches), presence or absence of eraser, and sharpened or unsharpened. The pencils were arrayed in a disorderly fashion on a table in front of *S*. While these three sets of materials involved different stimulus dimensions and different "story lines" were employed with them, an identical problem-solving task could be created with each set of materials.

Part-Task Training

Three programs, one for each of the three sets of materials, were developed to teach children to apply the technique of varying each factor in succession while holding all others constant. These programs had their origins in a free wheeling, loosely programmed training procedure which, nonetheless, achieved considerable success with bright first graders (Anderson, 1965). A modification of this procedure, which more nearly resembled a program, was first developed for the card array. After a cycle of tryout and revision the program was used with 10 second graders from a school in a rural community. On the last 10 frames of the program, upon which terminal behavior was required, seven children made no errors, one made one error, one made two errors, and one made three errors. The mean time for completion was 73 minutes.

The pencil program consisted of a literal translation of the card array program. In each frame, a pencil word or symbol was substituted for every card array word or symbol. No other modifications were made. Prior to the experiment, the pencil program was run with 10 naïve second graders from a school in a rural community. The mean time to complete the program was 104 minutes. On the last 10 frames, six children made no errors, three made one error, and one made two errors. The cowboy program was also created by a literal translation of the card array program; however the cowboy program was not used with any children prior to the experiment.

The final form of the programs embodied an analysis of the total problem-solving strategy into two major subskills. The first section of each program was designed to teach appropriate conclusion-drawing behavior. The Experimenter (*E*) began by naming concepts while *S* pointed to all of the positive instances of the concepts. Then the roles were reversed; *E* pointed to all of the positive instances of concepts and *S* named the concepts. Next, *E* pointed to sets of instances, some positive and some negative, in such a way that each set defined a concept; *S* named the concepts. For example, *E* might point to a long sharpened pencil with no eraser and a short sharpened pencil with no eraser, indicating that each of these "showed the secret," then a long sharpened pencil with an eraser and a long unsharpened pencil without an eraser, indicating that the latter two did not "show the secret." If the child responded "the secret is sharpened with no eraser," he answered correctly. When *S* could correctly name seven out of eight consecutive concepts given a set of defining instances (a criterion he was required to meet before proceeding), he was then judged to have acquired a satisfactory approximation of the conclusion-drawing skill.

The second component in the total problem-solving skill is the skill of selecting appropriate instances. To begin the section of the program teaching this skill, *E* pointed to an instance. The *S* was required to pick an instance which was different from *E*'s instance in a specified way but the same in every other way. For example, *S* might be instructed to "pick a pencil just the same as mine except that it is a different length." After several frames involving the stimulus dimensions of a task taken one at a time, *S* was then required to pick three instances, each of which differed in exactly one respect from the instance designated by *E*. When *S* reached a criterion of seven out of eight consecutive correct selections of sets of three instances, he had mastered the skill of selecting instances.

The final section of each version of the part-task program taught the child to integrate the conclusion-drawing skill with the instance-selection skill. In this section, the child selected a set of instances; using the child's instance, *E* defined a concept; finally, the child named the concept. Next, *E* began indicating whether each instance was positive or negative as soon as the child selected it, instead of waiting until he had selected the entire set. This latter procedure was the same as with the terminal problems, except that when the child selected an inappropriate instance he was corrected before being told whether the instance was positive or negative. The child was not corrected on the 20 terminal frames, each of which entailed a problem presented using the same procedures as were used for test problems.

A standard correction procedure, not expressly described within the program, was implemented by *E* whenever *S* made an error. The *E* created a new problem similar to the one upon which the error was made, told *S* the answer to the new problem, and then presented the original problem a second time. This procedure almost always prompted the correct response.

Whole-Task Training

Whole-task programs were developed for the card array, the cowboy game, and the pencil collection. The three versions of the program were equivalent in the sense that a systematic substitution of words and symbols would permit the literal translation of one version into another. The first section of the whole-task program was identical to the first section of the part-task program.

In this section E named concepts and S pointed to all of the positive instances of the concepts. Thereafter, S received terminal problems. The E designated a focus instance. The S's task was to choose instances until he could name the concept. Whenever S chose an instance E indicated whether the instance "showed the secret." With one exception, the procedures for presenting problems within the whole-training program were the same as the procedures for administering test problems to be described in the next section. The exception was that when S selected six instances during a training problem without solving the problem E told the child the correct concept. Feedback of this sort was not given during the last 20 training problems nor during test problems.

The part-task and whole-task programs were equated in terms of the total number of task-relevant overt responses required under the assumption of error-free performance. This measure resembles measures such as number of trials that can be applied to simple tasks. For example, a child who behaves ideally on a terminal problem will select three instances and state a conclusion, a total of four distinguishable overt responses. Each of the versions of the part-task program required a total of 360 task-relevant, overt responses whereas each of the versions of the whole-task program required 364 such responses. The first 28 and the last 80 responses (20 terminal problems) were the same for both programs. In between, those who received the part-task program were led to make a progression of 252 responses designed to teach them a conclusion-drawing skill, an instance-selection skill, and to integrate the two, as detailed earlier. The middle section of the whole-task program, on the other hand, contained 64 terminal problems which could have been solved with 256 overt responses. It should be emphasized that these calculations are based on the assumption of error-free performance. Of course, errors were made. Based on data collected during the experiment, the typical S who received the part-task program made an estimated 410 overt, task-relevant responses while the typical S who received whole-task training made about 620 such responses.

Procedure

The training and the test problems were presented by three female graduate assistants, each of whom had had 15 or more hours experience training and testing children prior to the experiment. The author monitored 1–2 hours of each E's preexperimental training and testing performance. Several staff conferences were held in which the letter and spirit of the procedures were detailed, ambiguities resolved, and difficult problems discussed. In addition, each E had a 10-page manual giving an overview of the experiment and summarizing training procedures, and a 7-page manual setting forth the procedures for administering test problems.

Each child was trained and tested by a single E. One-third of the Ss under each treatment in the experiment were run by each E. Training and testing sessions were scheduled to be 20 minutes in length. Unless the child was sick or some other circumstances such as a special school program intervened, the child received three sessions a week until he completed the training and the testing. Most sessions were conducted at three widely separated stations in a large general-purpose room in the cooperating elementary school.

For both part-task training and whole-task training there was a mimeographed copy of the program for each child. The child did not read the program. Rather, the program was a script that guided the behavior of E. Except as otherwise indicated, E adhered closely to the program, which described the stimulus S was to see, contained the verbatum language E was to use, and indicated the response or responses S was to give.

Under both training methods E made generous use of social reinforcement. The frequency and contingencies of reinforcement were not expressly indicated within the programs, but instead were under the extemporaneous control of E. Particularly with respect to the whole training procedure, which was quite aversive for some Ss (at the beginning of training, especially) E was coached to maintain a pleasant, nonjudgmental posture in the face of poor performance, and to find every opportunity to reinforce. Overall, E probably gave supplementary social reinforcement (in addition to feedback) for about every third correct response or chain of correct responses, except when S was doing poorly, in which case every correct response was reinforced.

Presentation and scoring of test problems

The procedure for presenting problems was illustrated with the cowboy materials on a preceding page. The E presented a "focus instance," that is, an exemplar of the concept. The S then selected instances until he could name the concept or until he had selected six instances without being able to name the concept, at which point the problem terminated. Each time S pointed to an instance, E told him whether the instance showed the concept (positive instance) or did not show the concept (negative instance). If S stopped trying to solve the problem or stated an incorrect conclusion with which he was apparently satisfied, E attempted to keep him performing with one of a series of standardized prompts. For example, if S made no task-relevant responses for a period of 10–15 seconds, E said "What are you going to do now?" If S stated an incorrect concept and made no further task-relevant responses for 10–15 seconds, E said "Are you sure that's my secret?" When these prompts failed to elicit further behavior, E presented a stronger prompt. If S accumulated enough evidence to solve the problem but did not volunteer a conclusion, E said "As soon as you are sure you know, tell me the secret." The language of all the prompts and the contin-

gencies for their use were prescribed. The E was permitted no unstandardized remarks. No feedback was given as to the correctness of responses to test problems, except as was indirectly involved in procedures already described.

At the beginning of each experimental session in which test problems were presented, S first solved a series of simpler problems designed to provide warmup, to furnish a high initial frequency of reinforcement, and to make sure that the concepts that S would subsequently have to attain were in his repertoire. In this orientation exercise, E named concepts while S pointed to all of the positive instances of each concept. There were 16 concepts treated in this manner including the 8 which would have to be attained in order to solve the problems later in the session. If S made an error he was prompted to make the correct response and, in addition, any item upon which an error occurred was repeated after an interval until S made an unprompted correct response.

The E wrote a protocol for each test problem, recording in coded form the instances S selected, the statements S made, and the statements E made in the sequence in which these occurred. Immediately following a day's administration of problems, Es exchanged protocols to check them for legibility, completeness, accuracy of subject and problem identification, and accuracy of coding.

An S solved a problem when he selected a series of instances and stated a conclusion such that the instances implied the conclusion and no other. Performance on terminal problems was scored on a 3-point scale as follows: (*a*) S solves the problem and neither makes any logically unnecessary choices of instances nor states any incorrect conclusions (2 points); (*b*) S solves the problem but makes one or more unnecessary choices or states one or more incorrect conclusions (1 point); (*c*) S fails to solve the problem (0 points).

The protocols were punched on cards and then scored on an IBM 1620 computer using a program written for this purpose. One of the virtues of mechanical processing was to provide a means for tracking down and eliminating the clerical errors which inevitably arise when a large quantity of hand-written material must be analyzed. An elaborate series of internal consistency checks was built into the program. The computer rejected 21 protocols, or 2.5% of the total number of protocols, because of coding errors. A keypunch error had been made in 11 cases. In the remaining 10 cases the fault was in the protocol. These protocols were examined and decisions were made as to what the codes should have been. Herein lies the only aspect of the analysis in which subjective judgment was required.

The terminal problems included within the training programs were scored on the same 3-point scale as the test problems. However, problems presented during training were scored on the spot by E and only an abbreviated protocol was written.

Design and Subjects

There were three treatment conditions. One group (P) received part-task training. Another group (W) received whole-task training. A control group (C) received no treatment. Every S in the former two groups received training with two sets of materials. One-sixth of the Ss in each of the training groups was assigned to one of the six possible permutations of the three sets of training materials taken two at a time. About 48 hours after completing training, Ss in the training groups received eight test problems to assess retention. The retention problems involved the second set of materials with which S received training. Each control S received the retention problems during his first experimental session. One-third of the control Ss received problems involving each of the three sets of materials.

About 48 hours after receiving the retention problems Ss received a series of eight test problems to assess transfer of training. The transfer problems entailed the set of materials which S had not encountered during training. One-third of the control Ss received transfer problems involving each of the three sets of materials, a different set than was encountered during the retention problems.

With respect to both the retention and transfer problems, S received 2 zero-dimensional problems, 2 one-dimensional problems, 2 two-dimensional problems, and 2 three-dimensional problems. The order of presentation of problems was randomized for each S independently of other Ss.

The Ss were 53 second-semester first graders from a predominantly middle-class school located in a new housing development on the outskirts of a Midwestern city of 30,000. These Ss were randomly selected from among all of the first graders in the school and randomly assigned to experimental conditions. Since the study was conducted over a 3-month period, the experimental conditions were scheduled in a predetermined random order. There were 18 Ss in Group W and in Group C but only 17 in Group P. There was to have been an eighteenth S in this group; however, the last S to be run (who appeared to be making normal progress) had to be dropped because of the impending end of the school year. He was replaced by a dummy case at the cell mean to balance the design for statistical purposes.

Classroom teachers administered the California Test of Mental Maturity (Long Form, 1963 Revision). Unfortunately one teacher found it necessary to terminate the examination in the middle of a subtest because of inattention and disorderly behavior, so it is not possible to report IQs or MAs. Raw score (not including Delayed Memory subtest score) means and standard deviations were 64.6 and 8.2 for Group P, 69.3 and 7.4 for Group W, and 66.4 and 7.0 for Group C ($F = 1.77$, $df = 2/52$, $p > .05$).

TABLE 1

MEAN TRAINING TIME IN MINUTES

Materials	First task		Second task	
	Part training	Whole training	Part training	Whole training
Cards	96.4	76.3	96.2	53.0
Cowboys	144.9	93.2	81.7	65.3
Pencils	135.9	98.2	107.7	85.2
All materials	125.7	89.2	95.2	67.8

Note.—The *SD*s (estimated from MS_e terms) were 25.9 for the first task and 24.4 for the second task.

RESULTS

Acquisition

Table 1 contains mean training times. Table 2 contains mean percentage of possible score on the last 12 training problems. Analysis of variance indicated that on both tasks Group P performed significantly better ($\alpha = .01$) on the problems whereas Group W completed training in a significantly shorter period of time. Based on the estimates of number of responses made during training, which were described earlier, and the training times that appear in Table 1, it is estimated that during training Group P made relevant, overt responses at the rate of about 3.7 per minute. The rate for Group W is estimated to have been about 7.9 per minute.

The mean training times for Group P were considerably higher than the times obtained during preexperimental development of the programs. Part of the discrepancy was no doubt due to the fact that second graders were used in most of the developmental work while first graders were employed in the experiment. Also at two points in the versions of the program used in the experiment *S* had to reach a criterion before proceeding. These criteria were not part of the preexperimental procedure.

It might be argued that if Group W had been allowed as much training time as Group P it would have performed as well on the terminal problems. Figure 1 pictures performance over blocks of 12 problems for Group W. Notice that performance reaches an asymptote by the sixth or seventh block on the first task. Consequently, it seems highly improbable that further practice would have improved the performance of Group W very much.

There were significant differences between materials on the first training task due in large part to the relatively poor performance with the cowboys. We have observed that the story line employed with the cowboys tends to interfere with the problem solving of some children, who insist that the "friends of the sheriff" must have rifles or must ride horses. Other investigators have made similar observations (Bruner et al., 1956, p. 111). Evidently by the time he reached the second task *S* had learned enough so that he was not distracted by the story line.

There were significant materials effects on training time due to the pencils. The pencils were handled by *S* and shuffled by *E* after each problem. These manipulations took time.

TABLE 2

MEAN PERCENTAGE OF POSSIBLE SCORE ON THE LAST TWELVE TRAINING PROBLEMS

Materials	First task		Second task	
	Part training	Whole training	Part training	Whole training
Cards	88.9	56.9	63.9	57.6
Cowboys	49.3	54.2	85.4	42.4
Pencils	84.0	43.7	81.2	63.2
All materials	74.1	51.6	76.9	54.4

Note.—The *SD*s (estimated from MS_e terms) were 17.3 for the first task and 18.0 for the second task.

Retention and Transfer

Table 3 presents the means for the retention and transfer problems. In both cases there were significant differences among treatments. Comparisons ($\alpha = .01$) using the Newman-Keuls procedure indicated that on retention problems Group P was superior to both the other groups and Group W was superior to Group C. On the transfer problems Groups P and W were not significantly different but both were superior to Group C.

There were significant differences in the

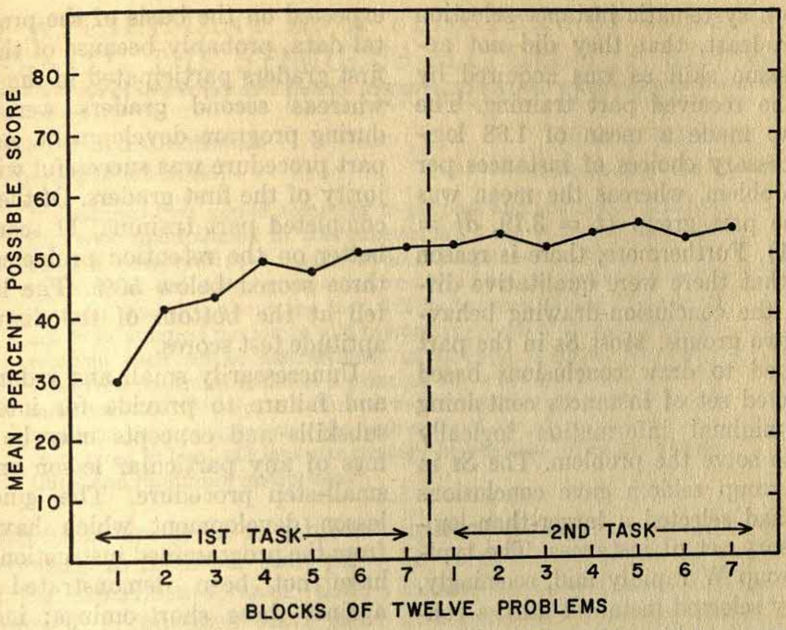


FIG. 1. Mean percentage of possible score on blocks of 12 training problems within the whole training group.

difficulty of the transfer problems according to the number of relevant stimulus dimensions the problems entailed. The means were 44.9, 49.1, 42.6, and 32.9% for the zero-, one-, two-, and three-dimensional problems, respectively. There was also a significant Materials × Dimensions interaction for the transfer problems, due primarily, for reasons which are not clear to the author, to the relatively great difficulty of the zero-dimensional card problems.

TABLE 3
MEAN PERCENTAGE OF POSSIBLE SCORE ON RETENTION AND TRANSFER PROBLEMS

Materials	Part training	Whole training	No training
Retention			
Cards	62.5	57.3	10.4
Cowboys	86.5	35.4	18.8
Pencils	76.1	55.2	16.7
All materials	75.0	49.3	15.3
Transfer			
Cards	44.8	58.3	30.2
Cowboys	49.0	41.7	21.9
Pencils	65.6	46.9	22.9
All materials	53.1	49.0	25.0

Note.—The *SD*s (estimated from MS_e terms) were 16.3 for the retention problems and 18.8 for the transfer problems when scores are pooled across dimensions.

DISCUSSION

The results indicate some limits to the generality of the rule proposed by Naylor and Briggs that whole training will be superior to part training for "highly organized" tasks. Of course it may be that there are characteristics of tasks, such as the amount and nature of its organization, which systematically interact with the length and complexity of the responses required from *S* at various stages during training, but this is evidently a matter about which there is much to be learned. The present author is pessimistic about the likelihood that broad generalizations concerning method-task interactions will emerge in the near future. Too much depends upon the specific features of the methods and the details of implementation of these features.

Both informal observation and the objective data suggest that in the present study *S*s who received whole training did

not acquire a systematic instance-selection skill, or, at least, that they did not acquire the same skill as was acquired by most *S*s who received part training. The whole group made a mean of 1.88 logically unnecessary choices of instances per retention problem, whereas the mean was 1.08 for the part group ($t = 3.19$, $df = 33$, $p < .01$). Furthermore, there is reason to believe that there were qualitative differences in the conclusion-drawing behavior of the two groups. Most *S*s in the part group learned to draw conclusions based on a restricted set of instances containing only the minimal information logically necessary to solve the problem. The *S*s in the whole group seldom gave conclusions until they had selected a larger-than-logically-necessary set of instances. The typical *S* in Group W rapidly and, seemingly, haphazardly selected instances until a conclusion occurred to him. Most *S*s in Group P, on the other hand, selected instances slowly and their behavior usually conformed to the method of varying each factor in succession while holding all other factors constant. At the point at which just enough information was available logically to solve the problem, the typical *S* who received part training usually offered a conclusion. The marked differences between Groups P and W in rate of response during training can be traced to the contrasting patterns of behavior typical of *S*s in the two groups.

It seemed possible that children who received part training would be able to solve a high percentage of the problems created with a new set of materials. Obviously this did not happen, indicating the need for a more refined method of producing generalized stimulus control (Anderson, 1965).

Group P did not do as well as would be expected on the basis of the preexperimental data, probably because of the fact that first graders participated in the experiment whereas second graders were employed during program development. Actually the part procedure was successful with the majority of the first graders. Of the 17 *S*s who completed part training, 11 scored 80% or better on the retention problems and only three scored below 50%. The latter three fell at the bottom of the distribution of aptitude test scores.

Unnecessarily small and redundant steps and failure to provide for integration of subskills and concepts may be shortcomings of any particular lesson employing a small-step procedure. The guidelines for lesson development which have emerged from the programmed instruction movement have not been demonstrated to guard against these shortcomings; indeed, it is possible that such deficiencies, particularly unnecessary redundancy, are endemic in currently available small-step, programed lessons. In the present instance, a small-step procedure worked relatively well. Whether a small-step, programed procedure would consistently prove best in other instances remains to be seen.

REFERENCES

ANDERSON, R. C. Can first graders learn an advanced problem-solving skill? *Journal of Educational Psychology*, 1965, **56**, 283–294.

BRUNER, J., GOODNOW, J., & AUSTIN, G. *A study of thinking*. New York: Wiley, 1956.

INHELDER, B., & PIAGET, J. *The growth of logical thinking*. New York: Basic Books, 1958.

NAYLOR, J. C., & BRIGGS, G. E. Effects of task complexity and task organization on the relative efficiency of part and whole training methods. *Journal of Experimental Psychology*, 1963, **65**, 217—224.

(Received May 30, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 3, 215-222

CONCEPT ATTAINMENT AS A FUNCTION OF INSTRUCTIONS CONCERNING THE STIMULUS MATERIAL, A STRATEGY, AND A PRINCIPLE FOR SECURING INFORMATION[1]

HERBERT J. KLAUSMEIER
University of Wisconsin

AND

DEAN L. MEINKE
Indiana State University

Instructions that varied according to purpose and amount of information were manipulated in this experiment. 1 set of instructions outlined a conservative focusing strategy, another set described the structure of the stimulus material, and a 3rd set presented only minimum information about the task. 102 Ss between 19 and 30 years of age were assigned randomly to 6 treatment groups of 17 each. 2 groups received each set of instructions; in addition, 1 group of each 2 was given information about a principle for securing information; the other was not. Concept attainment was more efficient for the 3 groups receiving the principle; also the rank order of the effects of instruction from most to least efficient was strategy instructions, structure instructions, and minimum instructions.

Verbal instructions are given to subjects (Ss) in experiments to facilitate their performance of the experimental task. Until recently, however, instructions have not been manipulated systematically to determine their effects. In part, lack of sufficient attention to the critical role of instructions in experiments is related to failure in specifying clearly the dimensions on which instructions may vary.

Instructions may vary according to purpose, method of presentation, amount of information presented, specificity of the information presented, and amount of nonverbal guidance. The latter three dimensions are relative and cannot be described precisely except in connection with a specific experiment. The method of presenting instructions, however, may be audio, visual, or audiovisual. Instructions may be formulated to achieve various purposes: (*a*) to acquaint *S* with the specific stimulus material or the more general task characteristics, (*b*) to acquaint *S* with the specific response or the more general performance desired, (*c*) to present *S* with information of a procedural type, such as a strategy, or a method, to apply to solution of the task, (*d*) to provide *S* with information of a substantive type, such as an advance organizer or a principle, to employ in performing the task, (*e*) to provide a set related to the recall or use of information or abilities, and (*f*) to manipulate the level of motivation of *S*.

Purposes *a* and *b* deal with clarification of the nature of the task and assume some degree of unfamiliarity of the task by *S*. Instructions used in experiments typically provide information concerning the nature of the stimulus, the response, or both. Purposes *c* and *d* involve the presentation of additional information, usually designed to facilitate performance of the task by *S*. Some research has been done with principles and advance organizers. Purposes *e* and *f* involve an attempt to directly manipulate thought processes or perceptions of *S*. For example, instructions may be designed to encourage *S* to recall relevant information or abilities that may be used in the present task. Also, instructions may be designed to produce varying amounts of stress in *Ss*, to indicate a reward system, to focus attention, or to produce other conditions assumed to be related to the level of motivation.

Research on instructions in experiments on concept learning is meager. The first experiment using instructions as an independent variable was by Archer, Bourne, and Brown (1955). They compared instructions encouraging analytic problem solving

[1] The research reported herein was performed pursuant to a contract with the United States Office of Education, Department of Health, Education and Welfare. This study is one of a series in concept learning of the Wisconsin Research and Development Center for Cognitive Learning under the direction of the senior author and part of the data gathered were included as part of the doctoral dissertation of the junior author.

to instructions encouraging nonanalytic problem solving. Analytic instructions did not significantly improve overall performance, but reduced variability and facilitated performance on the more complex tasks. Braley (1963) found no difference in efficiency of concept attainment between instructions oriented toward concept attainment procedures and those oriented toward problem solving. In contrast to the foregoing studies, Tagatz (1963) reported instructions and training with one type of stimulus material to have a negative effect upon performance with a second type of stimulus material. Thus, the effects of instructions on concept attainment are not clearly established. However, instructions are being used in many experiments on concept learning and are also of primary importance in curriculum research and development.

Instructions that convey information about a principle have a fairly consistent history of facilitating initial learning and transfer. Hilgard, Irvine, and Whipple (1953) had one group of *S*s learn to perform card tricks by memorizing a sequence while the other group learned a procedure for determining the sequence. Initial acquisition was faster for the memory group, and there were no differences in retention of the first trick. However, significantly more *S*s in the understanding group performed the second trick on the retention test. The understanding group, moreover, demonstrated significantly more transfer than the memory group to new tasks where simple transposition was not an effective aid to solution. Sassenrath (1959) found that learning to learn a principle during training facilitated learning to learn a reversal principle during the transfer period. Other experiments have also demonstrated the facilitating effects of the knowledge of a principle (Craig, 1956; Forgus & Schwartz, 1957; Haslerud & Meyers, 1958; Hendrickson & Schroeder, 1941; Judd, 1908). The results of many experiments are sufficiently clear to predict that providing information about a principle facilitates learning. In the present experiment, this prediction was tested with three sets of instructions that varied according to purpose.

The three sets of instructions developed for this experiment were varied deliberately according to purpose and indirectly in the amount of information provided. The method of presentation, specificity of information, and amount of nonverbal guidance were held constant. One set of instructions of about 400 words contained the minimum amount of information regarding the stimulus material and the desired responses necessary for the *S*s to proceed with the task. The second set of instructions of about 500 words incorporated the minimum instructions and more complete information regarding the organization, or structure, of the stimulus material. The third set of instructions of about 850 words incorporated the preceding and also a description of a conservative focusing strategy. The three sets of instructions and the instructions regarding the principle are included later in the method section of this article.

The prediction was that the strategy instructions would be associated with best performance, structure instructions next, and minimal instructions with least efficient performance. The prediction was based on the assumption that instructions could be written that would have a facilitative rather than an interfering effect. Also, the work of Bruner, Goodnow, and Austin (1956) suggests that an understanding of the relationship of the attributes and values incorporated in the stimulus material facilitates performance. In addition, the use of a conservative focusing strategy insures the attainment of the concept with greatest certainty that the concept identified by *S* as correct is in fact correct. Neither of these propositions regarding structure and strategy, however, has been tested experimentally. It may be observed also that the conservative focusing strategy could not be readily taught without an understanding of the structure of the material.

METHOD

Subjects

The *S*s were enrolled in educational psychology classes at the University of Wisconsin. They were assigned at random to each of six treatments with the restrictions of a proportional number of each

sex in each group, and an equal total number of Ss in each group. There were 102 Ss (90 females, 12 males) ranging in age from 19 to 30 years.

Experimental Material

The concepts were embedded in figural material on 128 stimulus cards, each 3 inches square. These cards were arranged in 16 rows and 8 columns on a large board. Each card contained seven attributes with two defining characteristics, or values, in all combinations as follows: border number, one or two; border continuity, solid or broken; figure number, one or two; figure size, large or small; figure texture, solid or spotted; figure color, red or green; and figure shape, circle or ellipse. The material is described in detail in Klausmeier, Harris, and Wiersma (1964). The concepts to be attained were conjunctive with three relevant attributes; for example, two large circles. Each S attained the same sequence of four concepts.

Experimental Procedure

The Ss were scheduled individually to come to the learning laboratory. After a brief introduction to the experimenter (*E*), the appropriate set of instructions was read to *S*. There were six sets of instructions: minimal with principle, minimal without principle, structure with principle, structure without principle, strategy with principle, and strategy without principle. Each *S* was allowed to attain as many concepts as possible, up to a maximum of 10, during a 55-minute period.

The minimal instructions presented only enough information for *S* to understand that he was to attain concepts. The structure instructions incorporated all the minimal information and also described the organization of the stimulus material according to the seven attributes, each having two defining characteristics as described previously. In addition to the description, the instructions required *S* to demonstrate that he could pick out at least one card representing each attribute and defining characteristic. The strategy instructions incorporated all the information of the minimal and structure instructions and also described a conservative focusing strategy for attaining concepts by selecting successive cards that differed from the focus card by only one defining characteristic. In addition, the instructions required *S* to demonstrate that he could pick out four successive cards that varied in only one defining characteristic from the focus card that was presented.

The general experimental procedure for all instructions was for *E* to present a focus card and for *S* to select successive cards as belonging to the concept of which the focus card was a member. The *S* could offer a hypothesis, his estimate of what the concept was, at any time after making the first card choice. No time limit was given *S*. He was instructed, however, to get the correct concept as quickly as possible.

The minimal instructions, the additional paragraph about the structure of the material, the two further paragraphs about the strategy, and the principle follow:

Minimal Instructions

This experiment is concerned with how people attain concepts. Your task is to attain several concepts that I have in mind. I am going to teach you how to attain the concepts. Your performance today in no way reflects on your intelligence. Also, there is nothing tricky about the experiment.

Now let us define what a concept is on this board. Concepts on this board are stated in terms of one or more of seven attributes listed on this slip of paper. For example, all the cards containing circular figures from the concept, circles. Show me four cards which belong to this concept. [The *E* waits until *S* points out four examples.] That's correct. Consider another example. All the cards with small red elliptical figures form the concept, small red ellipses. [Again *E* makes sure *S* knows four examples of yes-cards.) That's correct. A very large number of concepts can be formed, having one, two, or any combination of the seven attributes that are listed on your slip of paper. Please state a one-, then a two-, and then a three-attribute concept.... That's fine. In this experiment, your job is to attain concepts of the type we have just discussed. Do you have any question about what a concept is?

Listen carefully now to the procedure for attaining a concept that I have in mind. I shall indicate one card which belongs to the concept. This card we shall call the focus card. This focus card contains all seven attributes, and part of these seven attributes form the concept I have in mind. Your job is to test attributes of other cards in relation to the focus card to determine which attributes form the concept. Read off the number of the card you are checking and I shall tell you "yes" if it belongs to the concept and "no" if it does not belong to the concept.

When you think you know the concept, mark it on the slip of paper and give it to me. If the concept is correct, the task is completed. If not, I'll simply say "not correct, continue" and you will continue selecting cards until you again offer a concept. Use the slip of paper at all times and take it to the board with you.

Your job is to ascertain the correct concept as quickly as possible. Do you have any questions?

Structure Instructions

Now, let us examine the structure of the material. [The *E* points to the board.] There are 128 cards on this board. This slip of paper lists the attributes contained on each card on the board. Pick any one card on the board and point out the seven attributes, listed on your slip of paper, which are also on the card. Do your check-

ing aloud so that I may follow you. Start with number of borders. [The *E* checks that *S* follows the sequence given on the slip and checks all attributes.] Do you understand that any card you would have chosen would have all the seven attributes? Do you have any questions on the seven attributes contained on each card?

Strategy Instructions

The best way to ascertain the concept is to select your successive cards so that each card is exactly like the focus card, except for one attribute. Let us start with card #168 as an example of a focus card. To check the attribute, number of borders, we may go to card #61. Note that card #61 is exactly like the focus card except for the number of borders. To check the type of border we might go to card #95. Note that card #95 is exactly like the focus card except for type of border. Referring to your slip of paper, you tell me a card that differs by one attribute only for each of the five remaining attributes listed. Start with number of figures (6-no. of figures, 48-size, 88-texture, 105-color, 12-shape). You have just finished varying all seven attributes, one at a time, from the focus card. You should note that the cards which vary one attribute lie in the same row or column as the focus card. Do you have any questions about how to vary the seven attributes, one at a time, from the focus card?

If you had been trying to attain a concept of one or more attributes for which card #168 was the focus card, you would have proceeded in exactly the same manner. However, to some of your card choices I would have responded with "yes" and to others "no." On the basis of all the "yeses" and "nos" you could have determined the concept I had in mind.

The Principle

To get any concept I have in mind, you must use both the "yes" and "no" cards and there is an important principle to learn. The principle is that when you vary one attribute from the focus card and I give you a "no," the attribute of the focus card varied is part of the concept. When you vary one attribute from the focus card and I give you a "yes," the attribute is not part of the concept. For example, suppose #168 is the focus card and you select #61 which is like #168 except that it has two borders. If I said "no" to card #61 you would know that one border is part of the concept since the focus card is exactly like #61 except it has one border. However, if I said "yes" to #61 you would know that number of borders is not part of the concept since it makes no difference that #168 has one border and #61 has two borders. Let us try another example. Suppose #168 is the focus card and #12 is a "no" card. What does this tell you concerning the concept?...Circle is part of the concept since this is the only attribute varied from card #168.

This principle involving my giving you a "no" applies only when one attribute is varied from the focus card. Suppose I respond "no" to card #36. You could infer nothing about the concept. Can you tell me why? Correct. Because two attributes from the focus card were varied; size and texture. Do you have any questions about how to use the principle to use the 'no' cards to determine the concept?

Any question by *S* was answered by reading again the appropriate part of the instructions. When there were no further questions, the focus card for the first problem was presented. Time was kept from when *S* located the focus card until the correct concept was offered. The card choices and hypotheses offered by *S* were recorded serially. This permitted tallying the total number of card choices and the total number of hypotheses offered.

Experimental Design

A $2 \times 3 \times 4$ factorial design with repeated measures was used. The bileveled factor was prinicple versus no principle and the trileveled factor was instructions: minimal, structure, or strategy. The repeated measure was on each of four concepts attempted by each *S*. Consequently, there were 17 *S*s per cell with repeated measurements to ascertain the effect of ordinal position, or of practice across trials.

RESULTS

Since not all *S*s attained the same number of concepts, the analyses were carried out on the first four concepts attempted. In Table 1 are shown the means for each of the four concepts for the six groups according to principle or no principle and the type of instructions. From the data in Table 1, one can readily see that *S*s became more efficient with practice and that *S*s who received instructions incorporating the conservative strategy and the principle were the most efficient concept attainers.

TABLE 1

MEAN TIME IN SECONDS ON FOUR CONCEPTS

Treatment	Concept			
	A	B	C	D
Principle				
Strategy	378	265	107	106
Structure	513	352	206	145
Minimal	516	440	177	158
No principle				
Strategy	504	420	202	189
Structure	443	393	222	216
Minimal	523	506	297	255

The results of the analysis of variance using time to criterion as the dependent variable are shown in Table 2. The *F* ratio for the effect of a principle is significant beyond the .05 level. Similarly, the *F* ratio for the three types of instructions is significant beyond the .05 level. The effect for the ordinal position of the concepts, designated concepts in Table 2, is significant beyond the .01 level. These results indicate that instructions with a principle facilitate concept attainment and that *S*s improve performance with practice. This practice effect is, however, confounded with the difficulty level of the concepts. Since there is no basis for assuming that the concepts were of unequal difficulty, the improvement may be attributed to practice.

A second dependent variable was the number of card choices *S* made in attaining each of the four concepts. Table 3 shows the mean number of card choices for each of the four concepts for the six groups of *S*s according to principle or no principle and other instructions received. The fewest card choices were required by the groups receiving the strategy instructions, and the other two groups made about the same number of choices. The groups instructed about the principle required fewer choices than those not instructed.

The results of the analysis of variance using the number of card choices as the dependent variable are shown in Table 4. The results here are essentially the same as when time to criterion was the dependent variable. The effects of principle—no principle, of instructions according to the three purposes, and of the repeated measures of the concepts are significant beyond .01 level, indicating that the number of card choices was a more powerful measure than time to criterion. Apparently, the strategy instructions and the use of the principle either encouraged caution in making card choices or led to an increased amount of time to locate the specific cards that varied only one attribute from the focus card.

Table 5 shows the mean number of hypotheses offered by the six treatment groups on each of the four concepts. The fewest hypotheses were offered by the three groups

TABLE 2

ANALYSIS OF VARIANCE FOR MEAN TIME IN SECONDS

Source of variation	*df*	Mean square	*F* ratio
Grand mean	1	40,180,709	
Instructions (I)	2	261,892	3.76*
Principle (P)	1	461,638	6.63*
Concepts (C)	3	2,212,172	89.88**
I × P	2	86,460	*ns*
I × C	6	13,839	*ns*
P × C	3	24,330	*ns*
I × P × C	6	17,871	*ns*
Subjects within (I×P)	96	69,644	
Subjects (I×P) × C	288	24,612	
Total	408		

* $p < .05$.
** $p < .01$.

TABLE 3

MEAN NUMBER OF CARD CHOICES FOR FOUR CONCEPTS

Treatment	Concept A	B	C	D
Principle				
Strategy	10.5	9.7	7.1	8.1
Structure	18.8	19.5	15.7	11.9
Minimal	16.6	19.6	11.8	11.1
No principle				
Strategy	13.4	12.3	11.3	9.5
Structure	19.3	24.1	22.9	20.5
Minimal	26.0	27.0	18.9	20.1

TABLE 4

ANALYSIS OF VARIANCE FOR NUMBER OF CARD CHOICES

Source of variation	*df*	Mean square	*F* ratio
Grand mean	1	105,378	
Instructions (I)	2	3,474	12.42*
Principle (P)	1	2,971	10.62*
Concepts (C)	3	585	8.13*
I × P	2	252	*ns*
I × C	6	108	*ns*
P × C	3	26	*ns*
I × P × C	6	51	*ns*
Subjects within (I×P)	96	279	
Subjects (I×P) × C	288	72	
Total	408		

* $p < .01$.

TABLE 5

MEAN NUMBER OF HYPOTHESES OFFERED ON FOUR CONCEPTS

Treatment	Concept A	Concept B	Concept C	Concept D
Principle				
Strategy	1.3	1.3	1.0	1.2
Structure	2.1	2.2	1.4	1.2
Minimal	1.8	2.1	1.2	1.6
No principle				
Strategy	1.6	1.9	1.4	1.4
Structure	2.8	2.8	1.8	1.6
Minimal	2.2	3.1	2.0	1.6

receiving the principle. Fewest hypotheses prior to attaining the concepts were offered by the two groups receiving the strategy instructions. Those groups having the structure and minimum instructions offered about the same number of hypotheses.

Analysis of variance on the number of hypotheses to solution provides essentially the same results as for the preceding dependent variables. Effects beyond the .01 level of significance were obtained for instructions, the principle, and the concepts.

Finally, the results using the dependent variable of amount of potential information available at first hypothesis were considered. The amount of information each *S* could have potentially obtained at the time of offering a first hypothesis was determined by examining and comparing all his

TABLE 6

ANALYSIS OF VARIANCE FOR NUMBER OF HYPOTHESES OFFERED

Source of variation	*df*	Mean square	*F* ratio
Grand Mean	1	1,281	
Instructions (I)	2	15.5	8.31*
Principle (P)	1	25.0	13.37*
Concepts (C)	3	14.7	12.13*
I × P	2	0.5	*ns*
I × C	6	1.9	*ns*
P × C	3	1.4	*ns*
I × P × C	6	0.4	*ns*
Subjects within (I×P)	96	1.9	
Subjects (I×P) × C	288	1.2	
Total	408		

* $p < .01$.

TABLE 7

MEAN AMOUNT OF INFORMATION POTENTIALLY SECURED AT TIME SUBJECT OFFERED FIRST HYPOTHESIS ON FOUR CONCEPTS

Treatment	Concept A	Concept B	Concept C	Concept D
Principle				
Strategy	7.0	7.0	7.0	7.0
Structure	5.4	6.2	6.8	6.8
Minimal	5.5	5.9	6.6	6.5
No principle				
Strategy	6.2	5.8	6.9	6.8
Structure	4.5	5.3	6.2	6.3
Minimal	4.6	5.8	6.1	6.5

choices. The amount of information corresponds to the number of attributes tested and could range from one to seven. Table 7 shows the mean amount of information potentially obtained by each group. The strategy group with the principle obtained all the potential information necessary to attain the concept prior to stating a first hypothesis. No other group reached this level of proficiency.

Table 8 shows effects at the .01 level of significance for instructions, the principle, and the successive concepts. There was a significant Instructions × Concepts interaction. Apparently, this interaction is a reflection of poorer performance on the first two concepts by the structure group than the minimal group and a reversal of this

TABLE 8

ANALYSIS OF VARIANCE FOR AMOUNT OF INFORMATION POTENTIALLY SECURED AT FIRST HYPOTHESIS

Source of variation	*df*	Mean square	*F* ratio
Grand Mean	1	15,688	
Instructions (I)	2	26.8	11.44*
Principle (P)	1	30.8	13.14*
Concepts (C)	3	28.9	24.52*
I × P	2	1.0	*ns*
I × C	6	3.9	3.31*
P × C	3	1.6	*ns*
I × P × C	6	1.0	*ns*
Subjects within (I×P)	96	2.3	
Subjects (I×P) × C	288	1.2	
Total	408		

* $p < .01$.

trend on the third and fourth concepts, with greater gains accruing for the structure group than the minimal group.

DISCUSSION

As predicted, concept attainment was facilitated least by minimal instructions, next by instructions that presented information about the structure of the stimulus material, and most by instructions designed to teach a conservative focusing strategy. On all dependent variables, performance improved across trials, or concepts. One cannot determine experimentally whether the improvement was a function of practice or of the difficulty level of the concepts. Inasmuch as there is no basis for assuming unequal difficulty of the concepts and since the identical sequence of concepts was used with all *Ss*, one may assume that the improvement resulted from practice in attaining concepts of the conjunctive type. These results are important in terms of understanding, predicting, and controlling concept attainment in laboratory studies and also have implications for school learning of concepts.

The effect of practice on mean solution times was greatest from the second to the third concept. This might mean that positive transfer was maximized on the third concept. Clearly there was positive transfer from the first concept to successive concepts. In addition, the effect of interproblem transfer was lesser for instructions without the principle and greater for instructions with the principle. Thus, knowledge of the principle, especially how to use information provided by negative instances, was associated with consistently improved performance when time to criterion was used as the dependent measure.

Strategy instructions were closely associated with efficient performance in terms of number of card choices. With and without the principle, strategy instructions required the fewest card choices. It should be expected that the conservative focusing strategy might require more time to locate cards that varied in only one attribute from the focus card but that fewer card choices would be required to attain the concept. A further interesting observation is that knowledge of the principle facilitated performance more than knowledge of the structure. Apparently knowledge of the principle enabled *Ss* to secure more information from fewer card choices than did knowledge about the structure of the material. This may be interpreted as indicating that the instructions for the principle were better written or better understood by *Ss* than were those for structure. The intent was to write them equally well.

The combined effects of the conservative strategy and principle on hypothesizing behavior were dramatic. The *Ss* who had received these instructions potentially obtained all of the relevant information before offering an hypothesis, even on the first concept. No other group was as proficient, even on the fourth concept.

The importance of cognitive control of human behavior has been described by Miller, Galanter, and Pribram (1960). They stated that all human behavior is controlled by a plan, which in turn is comprised of strategies and tactics. Strategy refers to the control of the global units of behavior and tactics to the specific units. Bruner et al. (1956) reported the first extensive research with strategies in concept attainment and identified four strategies. Apparently the same person may use any of the four strategies in attaining concepts depending on situational conditions. The present experiment provides conclusive experimental evidence that not only a strategy but a principle for securing information efficiently can be described, and that providing this information to *Ss* markedly facilitates concept attainment over a short period of time.

The present experiment provides encouragement for moving ahead more rapidly on research in school settings in which attempts are made to develop and test instructions designed to facilitate concept learning in various subject fields. Even without further research it is safe to conclude that more time should be spent on teaching students how to learn concepts, the organization of the subject matter to be learned, and principles for securing and utilizing information. Further, educational

research on various methods of teaching has typically resulted in statistically nonsignificant results, whereas relatively small changes in instructions yield highly significant results in laboratory settings. Educational researchers apparently must find better means for controlling learning conditions in school settings, executing various treatments more systematically, and developing more sensitive measures of performance. Educational psychologists and other educational researchers have not demonstrated much inventiveness during the past decades in any of these matters.

REFERENCES

ARCHER, E. J., BOURNE, L. E., JR., & BROWN, F. G. concept identification as a function of irrelevant information and instructions. *Journal of Experimental Psychology,* 1955, **49,** 153–164.

BRALEY, L. S. Strategy selection and negative instances in concept learning. *Journal of Educational Psychology,* 1963, **54,** 154–159.

BRUNER, J. S., GOODNOW, J. J., & AUSTIN, G. A. *A study of thinking.* New York: Wiley, 1956.

CRAIG, R. C. Directed versus independent discovery of established relations. *Journal of Educational Psychology,* 1956, **47,** 223–234.

FORGUS, R. H., & SCHWARTZ, R. J. Efficient retention and transfer as affected by learning method. *Journal of Psychology,* 1957, **43,** 135–139.

HASLERUD, G. M., & MEYERS, S. The transfer value of given and individually derived principles. *Journal of Educational Psychology,* 1958, **49,** 293–298.

HENDRICKSON, G., & SCHROEDER, W. H. Transfer of training in learning to hit a submerged target. *Journal of Educational Psychology,* 1941, **32,** 205–213.

HILGARD, E. R., IRVINE, R. P., & WHIPPLE, J. E. Rote memorization, understanding, and transfer: An extension of Katona's card-trick experiments. *Journal of Experimental Psychology,* 1953, **46,** 288–292.

JUDD, C. H. The relation of special training to general intelligence. *Educational Review,* 1908, **36,** 28–42.

KLAUSMEIER, H. J., HARRIS, C. W., & WIERSMA, W. *Strategies of learning and efficiency of concept attainment by individuals and groups.* Cooperative Research Project No. 1442, University of Wisconsin, 1964.

MILLER, G. A., GALANTER, E., & PRIBRAM, K. H. *Plans and the structure of behavior.* New York: Holt, 1960.

SASSENRATH, J. M. Learning without awareness and transfer of learning sets. *Journal of Educational Psychology,* 1959, **50,** 205–212.

TAGATZ, G. *The effects of selected variables upon information processing and information utilization in concept attainment tasks.* Unpublished doctoral dissertation, University of Wisconsin, 1963.

(Received May 30, 1967)

EDUCATIONAL PSYCHOLOGIST

REQUIRED BY

PROVINCE OF NEW BRUNSWICK

DEPARTMENT OF EDUCATION

FREDERICTON

THE POSITION—This is professional educational psychology work in developing psychological testing services in the public school system within the province. The duties involve the responsibility for advising educational officials on the variety of tests, methods, and techniques used in the analysis and interpretation of psychological testing data to be used in various educational programs.

QUALIFICATIONS—University graduation with major course work in psychology supplemented by completion of courses in teaching methods and post graduate studies in educational psychology to the level of a Master's Degree and considerable teaching experience. Possession of a teacher's license and Certificate V as issued by the New Brunswick Department of Education and eligibility for registration under the Psychologist's Registration Act of the Province of New Brunswick are necessary requirements for this position.

SALARY RANGE—$8484 to $10,308 per annum

Competition number NB 68-60

APPLY—New Brunswick Civil Service Commission
P. O. Box 1055
FREDERICTON, New Brunswick

A new revision
of major significance . . .

APA PUBLICATION MANUAL
1967 Revision

The APA PUBLICATION MANUAL has undergone a complete revision for the first time in 10 years. Although the revised edition prescribes the same basic format for manuscripts as the previous edition, major changes and additions were made in the text in an effort to clarify and simplify the instructions on how manuscripts should be prepared.

The APA PUBLICATION MANUAL is intended primarily as a guide for authors and editors in the preparation of material to be published in APA journals. However, in preparing the original edition it was recognized that the MANUAL could, and should, be written to accommodate a wider range of manuscripts. The revised PUBLICATION MANUAL has retained that original concept.

$1.50

Order from:

AMERICAN PSYCHOLOGICAL ASSOCIATION
1200 Seventeenth Street, N. W.
Washington, D. C. 20036

Page 1

STATEMENT OF OWNERSHIP, MANAGEMENT AND CIRCULATION
(Act of October 23, 1962: Section 4369, Title 39, United States Code)

Publisher: File two copies of this form with your postmaster.
Postmaster: Complete verification on page 2

Form Approved, Budget Bureau No. 46-R029

1. DATE OF FILING

2. TITLE OF PUBLICATION
JOURNAL OF EDUCATIONAL PSYCHOLOGY

3. FREQUENCY OF ISSUE
Bi-monthly

4. LOCATION OF KNOWN OFFICE OF PUBLICATION *(Street, city, county, state, ZIP code)*
Mt. Royal & Gilford Sts. Baltimore, Maryland

5. LOCATION OF THE HEADQUARTERS OR GENERAL BUSINESS OFFICES OF THE PUBLISHERS *(Not printers)*
1200 17th Street, N.W., Washington, D.C. 20036

6. NAMES AND ADDRESSES OF PUBLISHER, EDITOR, AND MANAGING EDITOR

PUBLISHER *(Name and address)*
American Psychological Association 1200 17th Street N.W., Washington, D.C. 20036

EDITOR *(Name and address)*
Wayne H. Holtzman University of Texas Austin, Texas 78710

MANAGING EDITOR *(Name and address)*
Helen Orr 1200 17th Street, N.W. Washington, D.C. 20036

7. OWNER *(If owned by a corporation, its name and address must be stated and also immediately thereunder the names and addresses of stockholders owning or holding 1 percent or more of total amount of stock. If not owned by a corporation, the names and addresses of the individual owners must be given. If owned by a partnership or other unincorporated firm, its name and address, as well as that of each individual must be given.)*

NAME	ADDRESS
American Psychological Association	1200 17th Street N.W., Washington, D.C. 20036

8. KNOWN BONDHOLDERS, MORTGAGEES, AND OTHER SECURITY HOLDERS OWNING OR HOLDING 1 PERCENT OR MORE OF TOTAL AMOUNT OF BONDS, MORTGAGES OR OTHER SECURITIES *(If there are none, so state)*

NAME	ADDRESS

9. FOR COMPLETION BY NONPROFIT ORGANIZATIONS AUTHORIZED TO MAIL AT SPECIAL RATES *(Section 132.122, Postal Manual)*
(Check one)

The purpose, function, and nonprofit status of this organization and the exempt status for Federal income tax purposes
☒ Have not changed during preceding 12 months
☐ Have changed during preceding 12 months
(If changed, publisher must submit explanation of change with this statement.)

10. EXTENT AND NATURE OF CIRCULATION	AVERAGE NO. COPIES EACH ISSUE DURING PRECEDING 12 MONTHS	SINGLE ISSUE NEAREST TO FILING DATE
A. TOTAL NO. COPIES PRINTED *(Net Press Run)*	9,020	9,586
B. PAID CIRCULATION 1. SALES THROUGH DEALERS AND CARRIERS, STREET VENDORS AND COUNTER SALES	NONE	NONE
2. MAIL SUBSCRIPTIONS	7,812	8,027
C. TOTAL PAID CIRCULATION	7,812	8,027
D. FREE DISTRIBUTION *(including samples)* BY MAIL, CARRIER OR OTHER MEANS	86	85
E. TOTAL DISTRIBUTION *(Sum of C and D)*	7,898	8,112
F. OFFICE USE, LEFT-OVER, UNACCOUNTED, SPOILED AFTER PRINTING	1,122	1,474
G. TOTAL *(Sum of E & F—should equal net press run shown in A)*	9,020	9,586

I certify that the statements made by me above are correct and complete.
(Signature of editor, publisher, business manager, or owner)
Henry S. Salmon, Jr Bus Mgr

POD Form 3526 Aug. 1967

JOURNAL OF
EDUCATIONAL PSYCHOLOGY

Volume 59 August 1968 Number 4

CONTENTS

© 1968 by the American Psychological Association, Inc.

Journal of Educational Psychology
1968, Vol. 59, No. 4, 223-226

EFFECTIVENESS OF STUDY-SKILLS INSTRUCTION FOR HIGH SCHOOL SOPHOMORES[1]

WARREN L. HASLAM AND WILLIAM F. BROWN
Southwest Texas State College

LIBRARY

The Brown-Holtzman *Effective Study Course: High School Level* was taught to 74 high school sophomores during the fall of 1965 at Highlands High School, San Antonio, Texas. 59 of the 74 students receiving the instruction were individually matched with a control group of 59 students not receiving such instruction. Matching of the 2 groups was done on the basis of age, sex, race, intelligence quotient, subjects being studied, and 1st 9 weeks' grade-point average. Administration of the *Survey of Study Habits and Attitudes* before and after study-skills instruction indicated significant improvement in the measured study orientation of the experimental group. Following the course, students in the experimental and control groups were compared on 2 indexes of instructional results—9-weeks' course grades and scores on the *Effective Study Test*. The experimental group was found to be significantly higher on both indexes.

The assumption that improved study habits and improved academic effectiveness will result from systematic study-skills instruction is intuitively appealing. To date, however, published research on the productivity of study-skills instruction for high school students appears to be almost nonexistent. The investigation being reported was, therefore, undertaken to determine the effectiveness and acceptability of study-skills instruction for high school sophomores.

METHOD

Design

The research project was designed to determine if the Brown-Holtzman *Effective Study Course: High School Level* could produce significant improvement in the scholastic motivation, study behavior, and academic achievement of high school sophomores. Experimental students were selected and received instruction in how-to-study whereas an individually matched control group was denied such instruction. Upon completion of the course, the two matched groups were compared on two subsequent indexes of instructional results.

Subjects

In selecting experimental students, preference was given to those students who indicated on an application form a desire to continue their education beyond high school. Grades were also considered, with preference given to students with average grades; although some above and below average students were also selected to insure a representative cross-section of the student body. Finally, students having eight or more absences recorded during the previous school year were disqualified to help insure high attendance in the course. Thirty of the 65 applicants enrolled in fifth period study hall were accepted for the program; 44 of the 110 applicants for the after-school section were accepted.

Procedure

A mimeographed announcement describing the course objectives and content, the cost of materials, and the application procedure was distributed to all sophomore English classes and was posted on appropriate bulletin boards around the high school to inform students about the Effective Study Course. Sophomore students interested in obtaining more information about the course were directed to meet in a designated room either before or after school during the week of October 25. At each meeting, the instructor explained the course in more detail, answered the students' questions, and furnished interested students with application forms. The forms were taken home, completed, signed by either of their parents, and returned, together with a fee for materials.

Two sections of the course were offered—one during the instructor's normal conference period and the other after school. Sophomores enrolled in fifth period study hall were permitted to take the first, while all other applicants were required to take the section held after school.

To facilitate selection of a control sample, sophomore English teachers had their students complete an information form giving the student's name, age, sex, race, college plans, and subjects

[1] This study was supported, in part, by a grant from the Hogg Foundation for Mental Health, The University of Texas, Austin. It was based, in part, on an unpublished master's thesis by the senior author at Southwest Texas State College, 1966.

Copyright © 1968 by the American Psychological Association, Inc.

currently taken. Each student's grade average and intelligence quotient was obtained from the counseling office and added to his form. From the accumulated information, experimental and control students were individually matched on sex, race, age, intelligence quotients, first 9 weeks' grade averages, and subjects currently taken. Matching of the experimental group with a control group was done so that the two groups could be compared on two subsequent indexes of instructional effectiveness—third 9-weeks' course grades and scores on the *Effective Study Test* (Brown, 1964a). Of the original 74 students selected to receive the study-skills instruction, 59 students (experimental group) were finally matched with 59 students (control group) that did not receive study-skills instruction. The remaining 15 students were not included in the experimental group because it was not possible to match them individually with control students using the established limits of the matching criteria.

At the conclusion of instruction, each experimental student's reaction to instructor effectiveness, course content, and program acceptability was determined by administering a course evaluation questionnaire. To ascertain the unit's effect upon scholastic motivation and study behavior, the *Survey of Study Habits and Attitudes* (Brown & Holtzman, 1967) was administered to the experimental group both before and after instruction and the resulting scores were compared. Upon completion of the course, the *Effective Study Test* (Brown, 1964a) was administered to the experimental and control groups and test scores for the two groups were compared in order to determine their relative levels of study-skills knowledge. Finally, course grades for the experimental and control *S*s were collected and analyzed to determine the unit's influence upon subsequent scholastic success.

Course Description

The Brown-Holtzman *Effective Study Course: High School Level* is the result of a 14-year investigation into the factors determining scholastic success. The course is intended as a special instructional program for students prior to, or immediately following, high school entrance. The instructor's role is to serve as a program moderator or discussion leader rather than as a lecturer on the subject matter. The instructor is expected to encourage maximum student participation in the learning process to achieve the following course objectives: (*a*) to motivate each student toward developing more effective study habits; (*b*) to improve each student's study efficiency through better utilization of his study time; (*c*) to improve each student's study efficiency through improved organization of his study environment; (*d*) to improve each student's study efficiency through improved reading and writing techniques; (*e*) to improve each student's efficiency in preparing for and taking examinations; (*f*) to improve the self-direction of each student through the development of meaningful and realistic academic goals; and (*g*) to help each student develop a realistic understanding of high school life and peer acceptance problems.

Materials

A variety of instructional and guidance materials were used in the course. Three major items—the *Survey of Study Habits and Attitudes* (Brown & Holtzman, 1967), *Effective Study Guide* (Brown & Holtzman, 1964), and *Effective Study Test* (Brown, 1964a)—provided a systematic instructional program designed to help students improve their study skills. The *Survey of Study Habits and Attitudes*, hereafter referred to as SSHA, served as a motivating instrument; the *Effective Study Test*, hereafter referred to as EST, was used as an evaluative and instructional instrument; and the *Effective Study Guide* served as the students' textbook. Other instructional materials utilized in the course were the *Effective Study Workbook* (Brown, 1964b), *Reading and Remembering Guide* (Brown, 1962), *Study Skills Surveys* (Brown, 1965), *Daily Activity Schedule* (Brown, 1960), and *Student-to-Student Tips* (Brown, 1964d).

A special *Instructor's Manual* (Brown, 1964c) was employed to assist the teacher in planning the course content and using the instructional materials. The manual consisted of 20 detailed lesson plans, each designed for a 55-minute class period, and appended materials designed to facilitate and enrich instruction and discussion. The *Instructor's Manual* suggested that the lesson plans could be presented in whatever order the instructor deemed suitable for his situation; however, the 20 lesson plans were presented in the recommended order to insure an appropriate evaluation of the course.

RESULTS AND DISCUSSION

Pre-course and post-course administrations of the SSHA were employed to assess each experimental student's study orientation before and after instruction on how to study. The SSHA is an easily administered measure of study methods, motivation for studying, and certain attitudes toward scholastic activities important in the classroom. Since the SSHA yields separate study habits and study attitudes scores, its value lies in identifying, for each student, specific areas of deficient academic behavior and scholastic motivation which may handicap future school performance. Table 1 reports the means and standard deviations for pre-course and post-course scores on all seven SSHA scales. From this table it may be noted that each of the SSHA scales indicates a significant improvement in the study habits and atti-

tudes of the experimental group. Using Fisher's t test for correlated samples, differences between the pre-course and post-course means and standard deviations were found to be significantly higher on all seven SSHA scales following study-skills instruction.

The school administration felt that testing of the control students should be kept to a minimum so as not to interfere with their classroom instruction. Consequently, SSHA data were not collected on students in the control sample. Statistal data presented in the SSHA manual were, therefore, employed to help evaluate the course's effectiveness. In a reliability study for 237 ninth graders, the SSHA manual reports that the mean total score decreased 1.1 points and the standard deviation increased .3 points over a 4-week interval. By comparison, the means and standard deviations for the experimental group increased 43.3 and .7 points, respectively, over the same length of time. Comparison of the test-retest SSHA scores for these two samples clearly suggests that the results obtained for the experimental group should be attributed to the effectiveness of their study-skills instruction.

The EST was administered to the ex-

TABLE 1
COMPARISON OF PRE-COURSE AND POST-COURSE SURVEY OF STUDY HABITS AND ATTITUDES (SSHA) SCORES FOR THE EXPERIMENTAL GROUP

SSHA Scale	Pre-Course[a]		Post-Course[b]		M_D	t
	M	SD	M	SD		
Delay avoidance	21.3	8.5	34.2	8.4	12.9	13.84*
Work methods	21.9	7.8	34.5	8.2	12.6	12.90*
Study habits	43.2	14.7	68.7	15.5	25.5	14.90*
Teacher approval	29.7	8.5	39.4	6.0	9.7	10.12*
Educational acceptance	30.1	6.9	38.1	6.0	8.0	9.76*
Study attitudes	59.8	14.2	77.6	10.9	17.8	11.50*
Study orientation	103.0	26.5	146.3	25.8	43.3	14.59*

[a] Initial testing was accomplished on November 5, 1965.

[b] Retesting was accomplished on December 6, 1965.

* $p < .001$.

TABLE 2
COMPARISON OF EFFECTIVE STUDY TEST (EST) SCORES FOR THE EXPERIMENTAL AND CONTROL SAMPLES

EST Scale	Experimental Group[a]		Control Group[b]		M_D	t
	M	SD	M	SD		
Reality orientation	21.2	2.3	19.2	3.0	2.0	4.52*
Study organization	21.6	1.9	16.6	3.0	5.0	13.12*
Writing behavior	21.0	2.0	18.4	2.4	2.6	7.53*
Reading behavior	20.4	2.6	18.1	3.0	2.3	5.66*
Examination behavior	18.8	2.5	16.8	2.3	2.0	4.34*
Total study effectiveness	103.1	7.1	89.1	9.7	14.0	11.55*

[a] Testing was accomplished on December 2, 1965.

[b] Testing was accomplished on November 29–30, 1965.

* $p < .001$.

perimental and control groups upon completion of the how-to-study course in order to permit a comparison of their knowledge about efficient study practices. The EST is specifically designed to measure a student's knowledge about efficient study methods and the factors influencing their development. Table 2 reports the means and standard deviations of scores for the experimental and control groups on all six EST scales. From this table it is evident that the experimental group was significantly more knowledgeable about efficient study techniques. Using Fisher's t test for correlated samples, differences between the means and standard deviations for all six subscales were found to favor the experimental group at a significance level greater than .001.

The impact of study-skills instruction upon subsequent scholastic achievement was assessed by employing the grade-point averages for the first, second, and third 9 weeks' grading periods. Grade-point averages were calculated on the basis of 4, 3, 2, 1, and 0 points for letter grades of A, B, C, D, and F, respectively. The experimental students' grade average and standard deviation during the first 9 weeks'

grade period was 2.37 and .83, respectively. The control group's mean and standard deviation for the same time period was 2.33 and .84, respectively. During the third 9 weeks, the experimental sample's mean was 2.63 and the standard deviation was .90. The third 9 weeks' mean and standard deviation for the control students was 2.37 and .86, respectively. The total grade-point average increase for the experimental group during the third 9-weeks' period was .26 and at a level of significance greater than .001. During the same time period, the control groups' increase was .06 and at a significance level less than .10.

The experimental group's reaction to instructor effectiveness, course content, and program acceptability was evaluated upon completion of instruction by administering a specially constructed course evaluation questionnaire. Anonymous responses to the 60-item questionnaire were tabulated and converted to percentages. Tabulation of responses to all 60 statements indicated that the experimental students' reactions were decisively positive to all evaluated aspects of the how-to-study course.

In interpreting the research data, it should be kept in mind that motivation is an important factor influencing scholastic achievement. The motivation variable was not considered in matching the two groups of students, although motivation is partially implied in a student's grade-point average, and grade-point average was one criterion used for matching. Although the research method employed does have its limitations, the statistical analysis of data from the SSHA, EST, and grade-point averages generally indicates a level of significance acceptable to most authorities. One may conclude from the research data that the study-skills instruction given to the sample of high school sophomores did increase their knowledge about effective study procedures, did improve their overall study orientation, and did improve their subsequent academic achievement.

REFERENCES

BROWN, W. F. *Reading and remembering guide.* San Marcos, Texas: Effective Study Materials, 1962.

BROWN, W. F. *Daily activity schedule.* San Marcos, Texas: Effective Study Materials, 1960.

BROWN, W. F. *Effective study test: High school level.* San Marcos, Texas: Effective Study Materials, 1964. (a)

BROWN, W. F. *Effective study workbook.* San Marcos, Texas: Effective Study Materials, 1964. (b)

BROWN, W. F. *Instructor's manual for effective study course.* San Marcos, Texas: Effective Study Materials, 1964. (c)

BROWN, W. F. *Student-to-student tips.* San Marcos, Texas: Effective Study Materials, 1964. (d)

BROWN, W. F. *Study skills surveys.* San Marcos, Texas: Effective Study Materials, 1965.

BROWN, W. F., & HOLTZMAN, W. H. *Effective study guide.* San Marcos, Texas: Effective Study Materials, 1964.

BROWN, W. F., & HOLTZMAN, W. H. *Survey of study habits and attitudes Form S.* New York: Psychological Corporation, 1967.

(Received February 3, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 4, 227-232

LONG-TERM CORRELATES OF CHILDREN'S LEARNING AND PROBLEM-SOLVING BEHAVIOR[1]

GARY M. OLSON, LEON K. MILLER, GORDON A. HALE, AND HAROLD W. STEVENSON
University of Minnesota

46 girls and 51 boys enrolled in the 7th grade were presented a series of learning and problem-solving tasks including paired-associate learning, discrimination learning, incidental learning, concept of probability, conservation of volume, verbal memory, and anagrams. Only in incidental learning did the performance of girls exceed that of boys, but in all tasks except paired-associate learning and discrimination learning sex differences were found in the correlations between performance in the experimental tasks and school grades for 1 term in English, social studies, science, and mathematics. When IQ was partialed out there was a greater decrease in the number of significant correlations for girls than for boys. Performance in the initial block of trials in the 2 paired-associate tasks correlated significantly with school grades for boys, but not for girls. A repetition of the study with 73 additional Ss resulted in similar general relations, but some differences in the specific patterns of correlations.

Recent experimental studies of children's learning and problem solving have had little influence on educational psychology, except when they have been derived from the theory of Piaget (e.g., Freyberg, 1966). This is unfortunate, for there have been rapid advances during the past decade in our understanding of children's cognitive processes. The question is often raised whether the experimental tasks used by child psychologists have any relevance to the problems of classroom learning. Experimental studies usually involve short-term investigations of learning and problem solving in artificial, highly controlled situations, while classroom learning occurs over long periods of time and is concerned with complex materials and skills. There is no obvious reason to assume that the behavior sampled in the laboratory should be qualitatively different from that occurring in the classroom, but little attention has been paid to investigating this problem. Stake (1961) has reported the results of a factor analysis using school grades, scores on achievement, aptitude, and intelligence tests, and performance on learning tasks, and although he interprets his results in support of the definition of intelligence as the ability to learn, his data provide evidence for the predictive validity of learning tasks for school performance. A second study (Stevenson & Odom, 1965) investigated the relation between teacher's ratings of children's learning ability and children's performance in paired-associate and discrimination learning, concept formation, and anagram tasks. Highly significant correlations were found for two of the five tasks: paired associates and anagrams.

The present study provides additional information about the utility of learning and problem-solving tasks as predictors of school success. Seventh graders were given a series of tasks that were adapted from standard experimental learning and problem-solving tasks for group presentation. Performance on these tasks was correlated with school grades and the influence of intellectual level was determined by partialing out IQ scores. Additional informa-

[1] This study was supported in part by a grant (M-3519) from the National Institute of Mental Health, United States Public Health Service, to H. W. Stevenson and by Public Health Service Fellowships to G. A. Hale (MH-6668) and to L. K. Miller (MH-23721) from the National Institute of Mental Health. This study reports data from a larger project on learning and problem solving currently being conducted by the authors and Robert E. Klein. The writers express their appreciation to Daniel Bibelheimer, Miye Cohen, and the many other persons who assisted in this project. They also wish to thank the principals and teachers at the schools visited for their helpful cooperation in this study.

tion about children's school performance was obtained from teachers' ratings of the children's learning ability.

METHOD

Subjects

The subjects (*S*s) were 46 girls and 51 boys enrolled in three seventh grade classrooms of a junior high school in a middle-class area of Minneapolis. The average CA of the boys was 13.1 years (*SD* = .6), and of the girls, 13.0 years (*SD* = .4). The average IQ obtained from the verbal scale of the Lorge-Thorndike group test was 101.2 for the boys (*SD* = 13.3), and 104.2 (*SD* = 14.1) for the girls.

Grades

The *S*'s grades in four courses, English, social studies, science, and mathematics, were available for the quarter immediately preceding the study. The grades were coded so that an "F" was given a score of 1 and an "A+" was give a score of 13. The average grade was "C" (a score of 5.8), with a maximum difference of 1.7 points for the various courses. The standard deviations for the four courses varied from 2.9 to 4.2 for the boys and from 2.6 to 3.6 for the girls. For both sexes variability was smallest for English grades and greatest for science grades.

Tasks

This study reports data from eight learning and problem-solving tasks selected from a series of 13 tasks that were administered to *S*s as part of a larger study. The tasks that were eliminated were those that failed to discriminate school performance. They included a task involving the discrimination of abstract forms, one involving the discrimination of classes of common objects, probability learning, a test of the concepts of conjunction and disjunction, and a test of children's ability to estimate ages of adults.

All of the tasks were administered to intact classrooms by means of sound movies, some in black and white and some in color, depending upon the nature of the task. A narrator introduced each task, giving all necessary instructions. This method provided a careful control for the consistency of the instructions across classrooms. The *S*s responded in booklets that corresponded to the content of each film.

Each classroom was visited on 9 successive days, and the total time spent on a given day varied from 20 to 30 minutes. Different orders of presentation were used in the three classrooms so that each test was given at least once during the first 4 and once during the last 4 of the 9 days. The teacher introduced the experimenters (*E*s) as persons from the university who were interested in seeing how persons of *S*s' ages could perform on different kinds of tasks. The *S*s were told that their performance would have no influence on their school grades.

The following tasks were included:

Paired associates (abstract words). Six stimulus-response pairs were used. Stimulus elements were trigrams with high association values (e.g., DAG, KOT) and response elements were common abstract words (e.g., health, joy). The stimulus element of each pair appeared first on the screen, followed by the paired presentation of the stimulus and response elements. After the list had been presented, the first page of the response booklet appeared on the screen and the narrator gave the instructions for responding. The projector was stopped while *S*s circled the response element they thought was associated with each stimulus element. This procedure was followed for eight presentations of the list.

Paired associates (abstract forms). This task was identical to the preceding one except that different nonsense syllables were used for the stimulus elements and relatively simple Japanese characters were used as the response elements.

Discrimination learning. The *S*'s task was to discriminate common objects on the basis of their class membership. Four classes (i.e., tools, food, toys, people) were represented by four objects. In turn, each of the classes was associated with one of four geometric shapes. Four objects, one from each class, appeared across the bottom of each page of the booklet, with one of the geometric shapes centered at the top. On each trial the corresponding page of *S*'s booklet was shown on the screen. The *S*s were instructed to choose the object they thought was correct and after a short interval *E* pointed to the correct one on the screen. The correct pairings were arbitrary and differed across classrooms. A total of 64 trials was presented.

Incidental learning. An 8-minute skit with a simple plot was filmed in sound and color. The plot allowed the elaboration of incidental aspects of the film, such as dialogue, background actions, clothing, and set. The skit took place in a living room where a man and a woman were conversing about such matters as the content of the newspaper, the husband's tobacco, and similar everyday topics. Reference was made to the whereabouts of Andy, with the implication that he was the couple's child. A visiting woman and a deliveryman appeared as part of the plot, and gave further support to this implication. When Andy finally appeared, he turned out to be a dog.

No instructions were given before the film, and it was always shown on the last day of testing under the guise of a reward for participation in the study. When the film was over, *S*s were given a booklet containing 31 multiple-choice and true-false questions about the skit. Questions were asked about incidental aspects of the content, costuming, and physical setting of the skit.

Concept of probability. This film was in color. The narrator showed *S*s two boxes, one contain-

ing 30 red pegs and one with 30 white pegs. He then interchanged 10 pegs from each box, leaving two complementary 2:1 ratios of the different colors in the two boxes. Seven questions were asked about various probability relations one might expect if blindfolded one drew pegs from the boxes in various manners. For example, the narrator asked how many red pegs he would get if he drew three pegs from Box A, the assortment of pegs that would be produced with successive reaches into Box B, and whether Box A or B would be most likely to yield a red peg if only one could be drawn.

Conservation of volume. This task was a modification of the standard Piaget task. The film was in color, and the narrator presented two beakers of colored water, with several black lines indicating various levels on the beakers. Four questions were asked about the amount of water that would be displaced if a ball of clay were dropped in the second beaker after an identical ball had been dropped in the first. The first question used the ball of clay as it was, while the three remaining questions were asked about the level to which the water would rise when the clay had been modified in several ways: rolled into a sausage shape, squashed into a pancake, and cut into small pieces.

Verbal memory. This task was derived from the "Memory for Stories I: The School Concert" at Year X of Form M of the Revised Stanford-Binet. The narrator read the instructions, indicating *S*s would be asked questions about the story, and then read the one-paragraph story. The standard questions contained in this subtest were printed in *S*s' booklets. The maximum number of points was 14.

Anagrams. The *S*s were asked to make as many words as possible in 8 minutes from the letters in the word "generation." The narrator demonstrated the anagrams game with the word "federal," by constructing "flare," "lead," and "deer" as examples. Only those words found in the dictionary were allowed.

RESULTS AND DISCUSSION

The mean number of correct responses for each of the experimental tasks is presented in Table 1. None of the tasks appeared to be inappropriately easy or difficult for the seventh-grade *S*s, for the within-group variability was reasonably large for each of the tasks. In general, there were no significant sex differences in either level of performance or variability. In incidental learning, however, girls made a significantly greater number of correct responses than boys ($t = 3.28$, $df = 89$, $p < .01$). Despite the lack of significant sex differences in average level of performance, subsequent analyses were performed separately for boys and girls because of the possibility that sex differences might appear in the patterns of relations.

TABLE 1

MEAN NUMBER OF CORRECT RESPONSES FOR EACH TASK ACCORDING TO SEX OF SUBJECT

Task	Boys		Girls	
	M	SD	*M*	*SD*
PA (Abstract words)	34.86	11.76	38.76	9.56
PA (Abstract forms)	33.53	12.33	36.45	10.69
Discrimination learning	33.82	14.22	33.08	13.56
Incidental learning	21.04	3.79	23.37	2.97
Concept of probability	4.65	1.30	4.26	1.51
Conservation	3.23	1.15	3.41	.89
Verbal memory	7.86	3.83	8.80	3.19
Anagrams	18.68	7.81	22.02	8.08

Note.—*N* (Range) = 39–48 for boys, 39–46 for girls.

The significant correlations between performance on the experimental tasks and school grades are presented in Table 2. (The degrees of freedom in this and subsequent tables vary for different correlations depending upon the number of *S*s for whom information was available.) The most consistent correlations were found between grades and the two forms of paired-associate learning and discrimination learning. These correlations did not differ notably among the various courses.

There were pervasive sex differences in the types of relations found for the remaining tasks. The incidental learning and concept of probability tasks were significantly related to grades in all but one instance for boys, but in no case were they significantly related to girls' grades. On the other hand, the conservation task, verbal memory, and anagrams were consistently related to the grades of girls, but not of boys. There is no obvious basis for interpreting these differences. Nevertheless, the results offer clear evidence that the behavior sampled in brief experimental tasks is significantly related to long-term classroom performance. The processes involved in many of the tasks are apparently similar to those required for success in the classroom, even though the mode of presentation and the materials are highly dissimilar in the two situations.

The correlations of school grades with IQ and with performance on the experimental tasks are presented in Table 3. The correla-

TABLE 2

CORRELATION OF SCHOOL GRADES WITH PERFORMANCE ON EXPERIMENTAL TASKS

Task	English	Social studies	Science	Math
Boys				
PA (Abstract words)	.59**	.48**	.74**	.67**
PA (Abstract forms)	.63**	.60**	.50**	.61**
Discrimination learning	.35*	.38*		.36*
Incidental learning	.35*		.53*	.30*
Concept of probability	.40**	.37*	.61**	.41**
Conservation				
Verbal memory	.36*		.46*	
Anagrams	.45**			
Girls				
PA (Abstract words)	.33*	.51*	.52**	
PA (Abstract forms)	.47**	.60**	.56**	.37*
Discrimination learning	.49**	.41**	.45*	.47**
Incidental learning				
Concept of probability				
Conservation	.62**	.41*	.47*	.44**
Verbal memory	.65**	.58**	.69**	.43**
Anagrams	.61**	.70**	.47**	.47**

* $p < .05$.
** $p < .01$.

tions of grades and IQ were considerably higher for girls than for boys. Performance in the experimental tasks also tended to be highly correlated with IQ; however, there were no sex differences in the magnitude of the correlations.

To determine the relative influence of intellectual level on the relation between performance in the experimental tasks and school grades, correlations were computed between the latter two variables with the contribution of IQ partialed out. The results, summarized in Table 4, reveal striking differences for boys and girls. When these results are compared with those of Table 2, it can be seen that the number of significant correlations decreased from 21 to 13 for boys and from 23 to 5 for girls. Intellectual level thus tended to be a common determinant in the two situations of the performance of girls, but not of boys. After IQ was partialed out, the magnitude of the correlations remained high for boys, and in the conservation task became significant for science and mathematics. In three instances the correlations when IQ was partialed out were negative, twice for girls in social studies and once for boys in science. Performance of boys, both in the experimental tasks and in school, is apparently highly dependent upon factors other than intellectual level. The present data provide no indication of what such factors might be; however, it is likely that they involve differences in motivational and personality characteristics. Many studies have found, for example, that such characteristics as level of anxiety and achievement motivation tend to be more important determinants of the performance of males than of females.

TABLE 3

SIGNIFICANT CORRELATIONS BETWEEN LORGE-THORNDIKE IQ (VERBAL), SCHOOL GRADES, AND PERFORMANCE ON EXPERIMENTAL TASKS

Variable	Boys	Girls
School Grades		
English	.56**	.76**
Social studies	.48**	.82**
Science	.65**	.75**
Mathematics	.39**	.68**
Experimental Tasks		
PA (Abstract words)	.46**	.49**
PA (Abstract forms)	.37*	.53**
Discrimination learning		.40**
Incidental learning	.34*	
Concept of probability	.53**	
Conservation		.48**
Verbal memory	.56**	.61**
Anagrams	.61**	.60**

* $p < .05$.
** $p < .01$.

TABLE 4

SIGNIFICANT CORRELATIONS BETWEEN PERFORMANCE ON THE EXPERIMENTAL TASKS AND SCHOOL GRADES (IQ PARTIALED OUT)

Task	English	Social studies	Science	Math
Boys				
PA (Abstract words)	.45**	.34*	.65**	.60**
PA (Abstract forms)	.55**	.52**		.55**
Discrimination learning		.31*		
Incidental learning				
Concept of probability			−.54**	
Conservation			.61**	.47**
Verbal memory			.53**	
Anagrams				.32*
Girls				
PA (Abstract words)				
PA (Abstract forms)		.34*		
Discrimination learning				
Incidental learning			.44*	
Concept of probability				
Conservation				.31*
Verbal memory		−.40*		
Anagrams		−.51**		

* $p < .05$.
** $p < .01$.

TABLE 5

SIGNIFICANT CORRELATIONS BETWEEN PERFORMANCE ON FIRST TRIAL BLOCK OF EXPERIMENTAL TASKS AND SCHOOL GRADES

Task	English	Social studies	Science	Math
Boys				
PA (Abstract words)	.37*	.32*	.46*	.47**
PA (Abstract forms)	.49**	.58**	.57**	.52**
Discrimination learning				.34*
Girls				
PA (Abstract words)				
PA (Abstract forms)		.39*	.48**	
Discrimination learning			−.47**	

* $p < .05$.
** $p < .01$.

The sensitivity of the paired-associate and discrimination learning tasks as predictors of school grades was determined by correlating the mean number of correct responses during the first block of trials with grades. Only on these tasks could such correlations be computed, since the other tasks did not involve repetitive trials. The significant correlations are presented in Table 5. Performance on the first block of trials in the paired-associate tasks was remarkably sensitive in predicting the grades of boys. This is especially notable since the first block of trials represented less than a 3-minute sample of behavior. The factors producing the sex differences in the relations between early performance in the paired-associate tasks and grades are not clear. Whatever they are, however, was evident to the boys' teachers, for the teachers' ratings of effectiveness of learning were more highly related to performance on Trial Block 1 of the two paired-associate tasks for boys (r = .37 and .67) than for girls (r = .20 and .31).

Correlations were also computed between performance on the experimental tasks and scores on the Iowa Tests of Basic Skills. The patterns of relations were highly similar to those found for school grades, thus the data are not presented in detail. Again, approximately equal numbers of significant correlations were found for boys and girls, but when IQ was partialed out the decrease in the number of significant correlations was much greater for girls than for boys.

Each teacher was asked to divide the class into five approximately equal-sized groups according to the student's general learning ability. The correlations between these ratings and performance in the experimental tasks were significant for girls in all tasks except incidental learning and concept of probability. For boys, significant correlations were found for the two paired-associate tasks, incidental learning, and verbal memory.

TABLE 6

SIGNIFICANT CORRELATIONS BETWEEN GRADES IN GRADE 7 AND PERFORMANCE AT GRADE 6 WITH AND WITHOUT IQ PARTIALED OUT

Task	English	Social studies	Science	Math
Boys				
PA (Abstract words)	.65** (.55**)	.72** (.64**)		.36*
Incidental learning	.37*	.39*		
Concept of probability		(.42*)		
Conservation				
Verbal memory	.59**	.64** (.44*)		.44**
Anagrams	.65** (.43*)	.77** (.58**)	.73**	.59** (.46*)
Girls				
PA (Abstract words)	.69** (.62**)	.59** (.46*)		
Incidental learning				
Concept of probability	.60**	.53**	.51*	.52**
Conservation	.36*			
Verbal memory	.40*	.54**		.41*
Anagrams	.75** (.69**)	.62** (.44**)	.56**	

Note.—Values in parentheses are correlations between variables with IQ partialed out.
* $p < .05$.
** $p < .01$.

Additional information. Data were also available for 35 boys and 38 girls for whom course grades were obtained for the first quarter of their seventh grade at the same junior high school attended by the other *S*s in this study. (Grades are not given in the Minneapolis school system until the seventh grade.) These *S*s were given all of the tasks discussed earlier except paired associates (abstract forms) and discrimination learning when they were in the sixth grade, 7 months prior to the assignment of grades.

The significant correlations between performance in the experimental tasks and school grades are presented in Table 6. The correlations are of approximately the same magnitude as those found for the seventh-grade *S*s. The correlations between performance and grades with IQ partialed out are presented in parentheses in Table 6. Although the particular correlations that were significant differed somewhat from those found for the seventh-grade *S*s, the tendency again was for the proportion of significant correlations remaining after IQ was partialed out to be greater for boys than for girls. Correlations between performance on the first block of trials in paired associate (abstract words) and grades in the four seventh-grade courses were significant for boys ($r \geq .38$) except for science, but none of the correlations was significant for girls.

The study began with the general question of ascertaining whether the experimental tasks used by child psychologists in studies of learning and problem solving have any relation to the long-term learning that occurs in school. The answer to the question proved to be more complicated than was anticipated, for the predictive validity of many of the tasks differed, depending upon the sex of *S*. There was no tendency for the more complex, and possibly more school-like tasks to produce higher relations with grades than the simple rote-learning tasks. In fact, it is of interest that the two tasks that yielded the most consistently significant correlations, paired-associate learning and discrimination learning, were presented in the most highly structured context and used highly artificial materials.

Many questions remain unresolved in this preliminary study. The most important of these concern (*a*) the bases of the greater importance of intellectual factors for girls than for boys in producing significant correlations between performance in the experimental tasks and school grades, and (*b*) the stronger relation for boys than for girls between performance in the early phases of paired-associate learning and school grades. There is a surprising paucity of information helpful in answering such questions. Further investigations should, however, not only increase our understanding of the correlates of the learning process in children, but also might yield new and effective means of predicting children's academic success.

REFERENCES

FREYBERG, P. S. Concept development in Piagetian terms in relation to school attainment. *Journal of Educational Psychology,* 1966, **57**, 164–168.

STAKE, R. E. Learning parameters, aptitudes and achievements. *Psychometric Monographs,* 1961, No. 9.

STEVENSON, H. W., & ODOM, R. C. Inter-relationships in children's learning. *Child Development,* 1965, **36**, 7–19.

(Received May 15, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 4, 233–238

ROLE OF IRRELEVANT CUES IN THE FORMATION OF CONCEPTS BY LOWER-CLASS CHILDREN[1]

PHYLLIS A. KATZ
New York University

The present study sought to assess differences in concept formation abilities of children at varying developmental stages. The concept task employed was a perceptually oriented one, with stimulus dimensions verbalized in advance and positive exemplars of the concept continuously visible to the children. Results obtained on 1st-, 3rd-, and 5th-grade Negro lower-class males revealed differences in concept ability associated with both chronological age and IQ scores. Increasing the amount of irrelevant stimulus information elicited more errors in all age groups but this variable did not interact significantly with developmental level. The response latencies of the older and brightest Ss increased with the number of irrelevant cues, whereas those of the less intelligent children did not. This finding was interpreted as suggestive of possible developmental differences in information processing.

In recent years, there has been increasing theoretical and empirical interest in children's conceptual performances, and most investigations have reported positive relationships between age (chronological and mental) and degree of conceptual proficiency (e.g., Long, 1940; Long & Welch, 1941; Osler & Fivel, 1961; Piaget, 1930; Sigel, 1953). The direction of this developmental relationship is not invariant, however, as attested to by some recent work which demonstrated that older children may be inferior to younger children when more complex stimuli are employed (e.g., Friedman, 1965; Klugh & Roehl, 1965; Osler & Kofsky, 1965; Osler & Trautman, 1961).

Although it is clear that developmental variations exist in children's conceptual performance, the issue of what factors account for such differences has not elicited general agreement. Two theoretical positions have been particularly prominent in this area. The first position, advanced by Piaget and his coworkers (e.g., Inhelder & Piaget, 1958), stresses the role of unspecified maturational changes in underlying cognitive structures. A second position, postulated by Kendler and Kendler (1959), asserts that developmental differences in conceptual functioning are attributable to the advanced capacity for verbal mediation associated with higher maturational levels. Neither of these two viewpoints readily yields predictions of a negative relation between age and cognitive performance.

The present study attempted to assess an alternative explanation of developmental differences, derivable from an information theory approach to concept formation (Bourne & Restle, 1959; Hovland, 1952). This position appears capable of accounting for negative relationships previously obtained. Within this viewpoint, the parameter of primary concern is the informational value of the stimuli to the subject (*S*). Systematic addition of irrelevant stimulus information increases the difficulty of concept problems for adults, in terms of the number of errors and time required for solution (Archer, Bourne, & Brown, 1955). The possibility suggests itself that the same relation between stimulus complexity and concept difficulty exists in older but not in younger children because of a differential capacity for processing complex stimulus information. Thus, older children may take longer than younger children to solve certain kinds of conceptual problems, and may actually find them more difficult than younger children who

[1] This study was supported by Cooperative Education Grant No. 1099, SAE 9176, Office of Education, United States Department of Health, Education and Welfare. Thanks are extended to Barry Karp for his invaluable assistance in data collection and analysis.

are not as aware of the complexity. The purpose of the present study was to assess this possibility.

The concept formation task employed in the present investigation was a perceptually oriented one which did not involve memory or learning. Positive instances of the concept were continuously visible to the child, and all stimulus dimensions were verbalized prior to the concept task. These procedures were employed in order to minimize the role of variables other than conceptual ability which might contribute to age variation, such as stimulus familiarity, availability of labels and differences in recall.

The expectation in the present study was that developmental differences in conceptual performance would be obtained even when such possible confounding variables were controlled, and, furthermore, that information processing strategies should differ with developmental level. More specifically, it is predicted that the older and more intelligent the child, the more closely his performance should approximate that of adults. In accordance with Bourne and Restle's model (1959), it was predicted that task difficulty (as measured by both the number of errors and the latency of response) should show a more consistent relation to the amount of irrelevant stimulus information for the high developmental *S*s.

METHOD

Subjects

The *S*s were 72 Negro male children drawn from the first-, third-, and fifth-grade classes of two elementary schools in the Harlem area of New York City. The mean chronological ages of these groups were 6 years 10 months, 9 years 2 months, and 11 years 2 months, respectively, with corresponding standard deviations of 3.5, 6.5 and 5.9 months. These children had all participated in a previous study concerned with perceptual correlates of reading disability (cf. Katz and Deutsch, 1963), and the experimenters were, therefore, familiar to all *S*s. The children were primarily from lower socioeconomic backgrounds as assessed by parental questionnaires. Although this type of group has been characterized by a relatively high incidence of academic failure, there appears to be a paucity of data available on the conceptual abilities of such children, which was the major reason for the choice of these *S*s.

At each grade level, *S*s were dichotomized into high and low intelligence groups on the basis of Lorge-Thorndike IQ Test Scores (Lorge & Thorndike, 1959) and children were randomly chosen from these groups. The mean IQs of the high and low groups were 108.5 and 83, respectively.

The response latency scores of these children were compared with those obtained from a group of eight psychology graduate students with a mean chronological age of 24.3. Although the task was obviously very simple for this latter group, it was felt that there response latency patterns would provide an interesting point of comparison.

Stimulus Materials

The stimuli on which the concepts were based were geometric shapes which differed along four dimensions, each with three easily discriminable attributes that is, form (circle, square, or equilateral triangle), color (red, blue, or yellow), height (½ inch, 1 inch, or 1½ inches), and number (one, two, or three). These stimuli were pretested on comparable groups of kindergarten children to assess the discriminability of the concept exemplars. The shapes were cut from gummed paper and centrally placed on 3 × 5 inch white index cards. These index cards were then pasted into two loose-leafed booklets made of black construction paper, three to a page. One booklet contained three positive instances of each concept. The other booklet contained three choices for each concept. The *S*'s task was to point to the choice that was like all the positive instances. There was, of course, only one correct choice for each problem, and the position of this correct choice was randomly varied among the left, middle, and right index cards.

The concepts employed varied in two ways: (*a*) the type of cue relevant to solution (color, form, number, or size), and (*b*) the number of irrelevant stimulus cues (one, two, or three). In accordance with Archer, Bourne, and Brown (1955), an irrelevant cue was defined as one which varied in the positive exemplars of a concept, but was irrelevant to the correct solution. Thus, *S*s received 12 concept problems, each with one cue relevant to solution. The first four problems contained one irrelevant cue, the second four had two irrelevant cues, and the last four had three irrelevant cues. Thus, there were three different levels of stimulus complexity. At each level of stimulus complexity, there was one concept based on form, number, size, and color. These were randomly varied at each complexity level. Order of presentation was the same for all *S*s, namely, four concepts each with one, two, and three irrelevant cues.

Procedure

Each *S* was tested individually in the school. Upon entering the experimental room, *S* was informed that he would be playing a game in which he was to pick out pictures that looked

like each other. Two index cards containing identical red circles were then presented to *S*, and he was given the following instructions:

Here are two pictures. Do they look like each other? Why? [An attempt was made to elicit correct verbalizations of both dimensions.] Here are some other pictures. [Three additional index cards were introduced]. Can you pick out one picture from here that is like these two?

If *S* made the correct choice, the experimenter (*E*) said "good" and asked the child "Why is this one like these two?" If the child was unable to supply the correct verbalization, *E* provided it. Four other examples of this type were then introduced which utilized the cues of shape, number, size, and color. Verbalizations were elicited following the child's choices, and verbal reinforcement (i.e., "good," "that's right") followed each correct response and verbalization. Most of the children were able to verbalize correctly the relevant dimension. When they could not, *E* supplied it, and asked *S* to repeat it.

Following this verbalization training, the booklets were introduced with the following instructions:

Now we're going to do some more. This time the pictures will be in these books. In this book, all the pictures on the page are alike in some way. Look at them very carefully to see in what way they are alike. Now I'm going to show you some other pictures in this book. I want you to point to the picture in this book [choices] that goes with these [positive instances]. Point to the one picture here that is like *all* of these.

Three sample items of this type were administered, using concepts with two relevant redundant stimulus cues and no irrelevant information. These were employed to insure that *S* understood the instructions. Following these examples, the twelve concept problems were introduced and *S* was told that *E* would no longer tell him if he were right or not, but would tell him how well he did at the end of the session. Both choice scores and latency scores were recorded for each *S*. Latency was measured by means of a stopwatch which was begun by *E* as soon as the positive exemplars were visible to *S*, and stopped when *S* made a pointing response.

RESULTS

Effect of Type of Relevant Cue

The total number of correct responses made to each of the four types of relevant stimulus cues was analyzed by means of a mixed three-way analysis of variance (Lindquist, 1953). The main effect of type of stimulus cue was not significant (F = 1.73, df = 3/198), nor did it significantly interact with age or IQ. The main effects of age and IQ were statistically significant at the .05 level (F = 3.47, df = 2/66 and F = 5.56, df = 1/66 respectively). The age differences indicate that more correct responses on all items are associated with the older children. The mean number correct for the first-, third-, and fifth-grade groups were 7.29, 7.67, and 8.41, respectively. As expected, the brighter children made more correct responses than their less intelligent peers, a mean of 8.28 as compared with 7.28.

TABLE 1
MEAN NUMBER OF CORRECT CHOICES OF EACH GROUP

Group	Number of Irrelevant Cues		
	One	Two	Three
First grade—low IQ	2.85	2.08	1.58
First grade—high IQ	3.67	2.75	1.50
Third grade—low IQ	3.17	2.25	1.58
Third grade—high IQ	3.83	2.67	1.83
Fifth grade—low IQ	3.75	2.85	1.75
Fifth grade—high IQ	3.75	2.92	1.92
Total	21.02	15.52	10.16

Effect of Stimulus Complexity

The mean number of correct concept choices associated with each level of stimulus complexity for the various groups are presented in Table 1.

A mixed three-way analysis of variance (Lindquist, 1953) conducted on these scores revealed the main effects of age (F = 3.42, df = 2/66), IQ (F = 6.33, df = 1/66); and level of stimulus complexity (F = 93.01, df = 2/132) to be statistically significant. The effects of age and IQ on the number of correct choices were described above, i.e., the older and brighter the child, the more correct responses. The significant F value associated with level of stimulus complexity indicates that the addition of irrelevant cues increased the difficulty of the concept problem for these children. None of the interaction effects was significant, and it can be seen from Table 1 that the amount of irrelevant stimulus information showed approximately the same relation to performance (as measured in

TABLE 2
MEAN RESPONSE LATENCIES OF EACH GROUP

Group	Number of Irrelevant Cues		
	One	Two	Three
First grade—low IQ	47.8	33.2	24.5
First grade—high IQ	27.0	40.4	29.2
Third grade—low IQ	27.5	34.2	32.2
Third grade—high IQ	21.7	30.1	29.8
Fifth grade—low IQ	26.6	30.7	29.8
Fifth grade—high IQ	21.5	27.8	39.7
Adult reference group	15.5	29.4	57.7

terms of number of correct responses) at all age levels tested.

A second measure obtained on all the children was the latency of response to each problem. The mean latencies of each group at the various levels of stimulus complexity are presented in Table 2. Also included in Table 2 are the mean latencies of an adult reference group of eight psychology graduate students, which was included for purposes of visual comparison. The scores of this latter group were not included in any of the statistical analyses presented (although individual F tests revealed that their pattern did differ significantly from all of the children's groups).

The latency scores of the first-, third-, and fifth-grade children were analyzed by means of a three-way analysis of variance which indicates that none of the main effects was statistically significant. Thus, total response time was not associated with age, IQ, or stimulus complexity. Stimulus complexity, however, did interact significantly with both age (F = 3.35, df = 4/132) and intelligence level (F = 4.25, df = 2/132). The specific direction of the Stimulus Complexity × Age interaction can be observed in Figure 1.

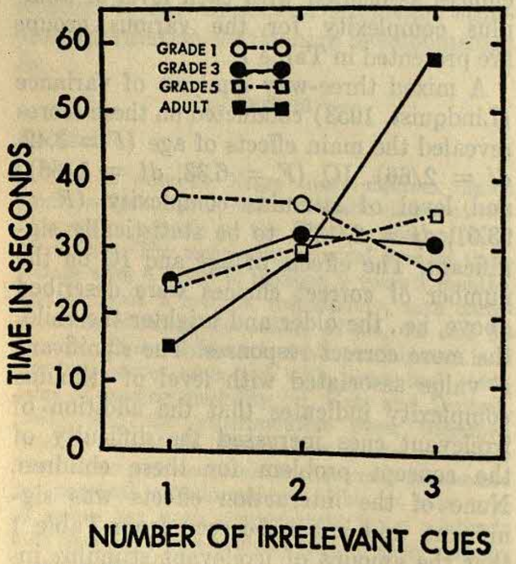


FIG. 1. Response latencies of each age group on concept problems of varying complexity.

It can be noted that the response times of the adult reference group on this task are directly proportional to the amount of irrelevant stimulus information contained in the concept problems. The various age groups tested, however, each exhibited differential patterns of response latencies. The fifth-grade children in the present study exhibited a pattern of responses similar to the adults, that is, their mean latencies increased with the addition of irrelevant cues. The youngest group, on the other hand, exhibited an opposite pattern with regard to latency measures. Response times of the first-grade group decreased as the problems became more difficult. The performance of the third-grade children was somewhere between these two extremes. They exhibited an increase in reaction time between the simpler (one irrelevant cue) and intermediate (two irrelevant cues) complexity problems, but not between the intermediate and difficult (three irrelevant cues) problems.

The means involved in the significant Stimulus Complexity × IQ interaction parallel the Stimulus Complexity × Age interaction. The more intelligent children at all age levels tended to approximate the adults' performance more closely. The low IQ children, on the other hand, exhibited a different pattern of response, one which was more similar to the youngest children tested.

DISCUSSION

The results of the present investigation indicate that proficiency in concept formation is related to developmental level in children, as assessed by both chronological

age and IQ. This finding is in accordance with a number of other studies (e.g., Inhelder & Piaget, 1958; Long & Welch, 1941; Osler & Fivel, 1961; Sigel, 1953). Unlike earlier studies, however, the present investigation attempted to minimize several other possible sources of age variation which may not be specifically related to conceptual capacity, such as memory, stimulus familiarity, and the availability of verbal labels. Toward this end, the conceptual task employed contained readily discriminable perceptual dimensions, with explicit verbalization, and continuously visible positive exemplars. The finding that developmental differences continue to manifest themselves in a task of this type supports the developmentalist view that gross qualitative differences in concept formation ability exist at varying maturational levels.

One of the interests of the present study was to assess possible differences in information processing strategies in children's conceptual performance. The present findings suggest that the parameter of irrelevant stimulus information is very significantly related to concept problem difficulty in children. The number of correct responses was inversely related to the level of stimulus complexity at each of the age levels tested. This finding is in accordance with results obtained with both college students (Archer et al., 1955) and with younger, somewhat more intelligent and higher social-class children in a concept-learning task (Osler & Kofsky, 1965). There was, however, a lack of interaction between stimulus complexity and developmental level on the number-of-correct-responses measure, which was not in accordance with expectation.

One finding that was suggestive of differential strategies was the significant interaction of stimulus complexity with both age and IQ with regard to response latencies. An exponential relationship between response latency and stimulus complexity was exhibited by the group of graduate students in the present study, a finding very similar to that obtained by Archer et al. (1955) with college students employing much more complex conceptual tasks. In general, the older and brighter the child, the more closely his reaction-time pattern approximated that of the adult group. Thus, the fifth-grade, high-IQ *S*s showed the clearest increase in reaction time as additional irrelevant cues were introduced. The first-grade children, on the other hand, exhibited an opposite pattern. Their response times decreased somewhat as the number of irrelevant cues increased. This trend was particularly pronounced in the youngest children with the lowest IQs.

Although other interpretations are possible, the differential relation between response latency and stimulus complexity at the various developmental levels assessed in the present investigation suggests that the older and more intelligent children may have a greater capacity for processing more stimulus information than children at less advanced cognitive stage. The work of Kagan (1965) and his associates on reflective and impulsive cognitive styles appears relevant to the present interpretation since, as Kagan suggests, it is possible that longer response times are indicative of a general reflective mode. It should be noted, however, that differences between age and IQ groups in overall reaction times were not obtained in the present investigation. Longer response latencies were exhibited by the more intelligent children only with regard to the more complex stimuli, thus suggesting that reflection was not a general response characteristic but rather one that was appropriately related to the stimulus characteristics of the task.

The type of relevant stimulus cue employed was not differentially related to the difficulty of the concept problem. This finding, although in accordance with adult studies (e.g., Archer et al., 1955), appears to run counter to a number of studies showing that certain stimulus cues may be more salient than others in the perceptual and cognitive processes of younger children (e.g., Brian & Goodenough, 1929; Corah, 1966; Kagan & Lemkin, 1961). Unlike these earlier studies, however, the procedure employed in the present study introduced the relevant stimulus dimensions together with appropriate labels to *S* prior to the con-

cept task. It would appear then that the perceptual trends observed in earlier findings may be easily changed by appropriate verbal instructions. In this regard, it is interesting to note that the absence of cue differences was obtained in the present investigation with a group of lower-class children who would be least likely to be very familiar with the type of stimuli employed. Another possible explanation which suggests itself for the discrepancy between present results and earlier findings is that the previously obtained dominance of certain stimulus cues in young children may have been reflecting differential availability of labels for these dimensions, rather than developmental differences in perceptual processes. Future investigation is indicated to assess the possibility.

REFERENCES

ARCHER, E. J., BOURNE, L. E., JR., & BROWN, F. G. Concept identification as a function of irrelevant information and instructions. *Journal of Experimental Psychology,* 1955, **49**, 153–164.

BOURNE, L. E., & RESTLE, F. Mathematical theory of concept identification. *Psychological Review,* 1959, **66**, 278–296.

BRIAN, C. R., & GOODENOUGH, F. L. The relative potency of color and form perception at various ages. *Journal of Experimental Psychology,* 1929, **12**, 197–213.

CORAH, N. L. The influence of some stimulus characteristics on color and form perception in nursery-school children. *Child Development,* 1966, **37**, 205–212.

FRIEDMAN, S. R. Developmental level and concept learning: Confirmation of an inverse relationship. *Psychonomic Science,* 1965, **2**, 34.

HOVLAND, C. I. A "communication analysis" of concept learning. *Psychological Review,* 1952, **59**, 461–472.

INHELDER, B., & PIAGET, J. *The growth of logical thinking from childhood to adolescence.* New York: Basic Books, 1958.

KAGAN, J. Reflection-impulsivity and reading ability in primary grade children. *Child Development,* 1965, **36**, 609–628.

KAGAN, J., & LEMKIN, J. Form, color and size in children's conceptual behavior. *Child Development,* 1961, **32**, 25–28.

KATZ, P., & DEUTSCH, M. Relation of auditory-visual shifting to reading achievement. *Perceptual and Motor Skills,* 1963, **17**, 327–332.

KENDLER, T. S., & KENDLER, H. H. Reversal and non-reversal shifts in kindergarten children. *Journal of Experimental Psychology,* 1959, **58**, 56–60.

KLUGH, E., & ROEHL, K. Developmental level and concept learning: Interaction of age and complexity. *Psychonomic Science,* 1965, **2**, 385–386.

LINDQUIST, E. F. *Design and analysis of experiments in psychology and education.* Boston: Houghton-Mifflin, 1953.

LONG, L. Conceptual relationships in children: The concept of roundness. *Journal of Genetic Psychol.,* 1940, **57**, 289–315.

LONG, L., & WELCH, L. Reasoning ability in young children. *Journal of Psychology,* 1941, **12**, 21–44.

LORGE, I., & THORNDIKE, P. I. *Lorge-Thorndike Tests of Intelligence,* Specimen Test Booklet. New York: Houghton-Mifflin, 1959.

OSLER, S. F., & FIVEL, M. W. Concept attainment: I. Effect of age and intelligence in concept attainment by induction. *Journal of Experimental Psychology,* 1961, **62**, 1–8.

OSLER, S. F., & KOFSKY, E. Stimulus uncertainty as a variable in the development of conceptual ability. *Journal of Experimental Child Psychology,* 1965, **2**, 264–279.

OSLER, S. F., & TRAUTMAN, G. E. Concept attainment: II. Effect of stimulus complexity upon concept attainment at two levels of intelligence. *Journal of Experimental Psychology,* 1961, **62**, 14–23.

PIAGET, J. *The child's conception of physical causality.* New York: Harcourt Brace, 1930.

SIGEL, I. The developmental trends in the abstraction ability of young elementary school children. *Child Development,* 1953, **24**, 131–144.

(Received May 18, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 4, 239–243

EFFECTS OF NOISE ON PUPIL PERFORMANCE[1]

BARBARA R. SLATER
Pelham Public Schools, Pelham, New York

263 7th-grade public school children were tested to determine whether quiet (45–55 db.), average (55–70 db.), and noisy (75–90 db.) classroom and experimental conditions had a relationship to written task performance of relatively short duration. It was hypothesized that Ss would perform better under quiet than under average and noisy conditions and that boys would be more detrimentally affected by noise than girls. Noise typical of that experienced in schools and white noise were both used. Means and standard deviations were compared across the conditions used and analyses of variance were performed on the data. No noise effect, either detrimental or facilitating, was demonstrated on speed or on accuracy of performance. Ss' perceptions of the effects of noise and measured anxiety had little relationship to actual performance.

Within the past several decades interest in improving educational facilities has seen a steady increase. One of the areas around which much controversy has arisen is that of acoustical environment in the schools. In order to delineate a clear-cut area for investigation, only the specific category of the effect of noise upon written task performance of children was treated in the present study.

Problem

The effects of noise upon human performance has been an area of conflicting reports and research studies for over four decades. Researchers such as Broadbent (1953, 1954, 1958a, 1958b), Grimaldi (1958), Jerison (1959), Kitamura (1964), Lehmann, Creswell, and Huffman (1965), and Weston and Adams (1935) have reported detrimental effects of noise upon performance. Other researchers, such as Super, Braasch, and Shay (1947), Park and Payne (1963), Sanders (1961), Teichner, Arees, and Reilly (1963), and Tinker (1925), have reported either equivocal results or no evidence of a detrimental noise effect.

Much of the existing research has used adult subjects (*S*s), artificial conditions, and/or noise levels exceeding those encountered in any but extreme situations. Planning of educational facilities based upon these studies, without evidence of comparability of *S*s or conditions may be somewhat misleading. The conflicting results of studies pertinent to educational application further indicate the need for additional research to assist in determining the need for and value of acoustical treatment in schools.

The major objective of this study was to investigate the effects of noise upon written performance in a realistic environmental setting while meeting as many of the requirements of noise research as possible. Such an application necessitated the use of an actual classroom environment, tasks pertinent to school routine, and noise comparable to that encountered by children during school activities. While it might be interesting to prove that noise of approximately 100 decibels transmitted, via earphones, to children working in isolation chambers caused a deterioration in performance, it would be rather difficult to generalize to children who are not subjected to this level of noise, who are not equipped with earphones, and who do not work in isolation chambers. It was also necessary to control for or to take into ac-

[1] This paper is based on a doctoral dissertation submitted to the Joint Committee on Graduate Instruction, Columbia University. Committee members were: Mary Alice White (Chairman), Stanley Lieberfreund, and Walter H. MacGinitie.

The research reported herein was performed pursuant to a contract with the United States Department of Health, Education, and Welfare, Office of Education, under the provisions of the Cooperative Research program.

The author was a graduate student in school psychology at Teachers College, Columbia University, at the time the research was conducted.

count physical variables, differences in set caused by instructions and task perceptions, individual differences, the so-called Hawthorne effect of a positive change resulting from any alteration of conditions, and the carry-over effect from one condition to another. The noise characteristics and instrumentation used had to be described for purposes of analysis and possible replication.

A secondary objective was to investigate individual differences under conditions of noise. According to Goodenough (1954, pp. 482–483), girls achieve slightly better in school, possibly because of girls' greater docility and better application to studies. Terman and Tyler (1954, pp. 1064–1114) cited studies indicating that boys may be more physically active and restless than girls. Considering this, it was expected that boys would be less motivated to work than girls and would be more prone to distraction under conditions of noise.

In line with Sarason's (Sarason, Davidson, Lighthall, Waite, & Ruebush, 1960; Sarason & Gordon, 1953) work on anxiety in children, it was decided to investigate the relationship between anxiety and the effects of noise. The *S*s' perceptions of noise and the effects of set toward noise and toward the experiment were also considered as possible secondary factors.

Noise was defined, consistent with Peterson and Gross (1963), as undesired sound. Since the normal noise encountered by children in school is intermittent or irregular-interval noise, the study used this type of noise.

Hypotheses

Hypothesis 1. Under conditions of irregular-interval noise varying approximately 75–90 decibels, children's task performance will be lower than under conditions of relative quiet of approximately 45–55 decibels.

Hypothesis 2. Under conditions of irregular-interval noise at levels of 75–90 decibels, children's task performance will be lower than under conditions of normal or average classroom noise of approximately 55–70 decibels.

Hypothesis 3. Under conditions of relative quiet, children's task performance will be higher than under conditions of normal classroom noise.

Hypothesis 4. Performance levels of boys will be lower than performance levels of girls under conditions of irregular-interval noise of 75–90 decibels.

Method

Subjects

The *S*s were 129 male and 134 female seventh-grade children from a centralized suburban school on the outskirts of a small urban complex, consisting of three cities, in south-central New York state. No children with hearing deficiencies were included.

Task

The STEP Reading Test, Form 3 was used as the written task. The *S*s were permitted to answer as many of the 70 questions as they were able during the 30 minute testing period in order to prevent a ceiling effect. Two experimental sets were assumed; the tension of a test situation and the more relaxed atmosphere of a homework situation. These assumed sets were induced through differences in answer sheets, instructions, and examiner behavior. Measures of speed, as the total number of questions attempted, and of accuracy, as the percentage correct of the number attempted, were obtained.

Grouping

Eight equated groups were used and each *S* was tested only once. Four days prior to the experiment, Part 1 of an alternate form of the STEP Reading Test was given to each *S*. The total number of correct answers on this pretest was used as the essential basis of equating the eight groups. Scores were arranged from high to low and a matching process ensured that the groups were comparable.

The groups were also roughly equated, by matching, on the basis of IQ, socioeconomic status and achievement to ensure that no systematic differences among groups on these variables occurred and to permit examination of individual differences. Each group consisted of approximately equal numbers of males and females.

After completion of the matching process, a testing condition was assigned randomly to each of the eight groups. The composition of these groups is presented in Table 1.

Noise Conditions

A study was conducted of possible methods of producing suitable noise conditions, as determined by past research and by surveys of typical noise levels found in schools, and of methods of meas-

uring and analyzing such noise. Sound-pressure levels in decibels were used throughout the study. Typical noise levels from the sample school are presented in Table 2. These levels are comparable to levels reported by Fitzroy and Reid (1963) in an extensive survey of 37 schools.

The hypothesized levels for the quiet (45–55 decibels), average (55–70 decibels), and noisy (75–90 decibels) conditions were selected so as to avoid exceeding the minimum and maximum limits which might be anticipated to occur within a school environment. Noise characteristics of the experiment were as follows:

Part 1: Classroom. The classroom section was divided into quiet, average, and noisy conditions. Only noise familiar in some degree to *Ss* was used and testing was done in classrooms with the usual row type seating.

For the quiet condition (45–55 decibels) the test room was maintained in an isolated situation. The surrounding classrooms were empty; the corridors were kept free from passage; and bells, buzzers, and intercom systems were temporarily eliminated.

For the average noise condition (55–70 decibels), testing was done while classes were being conducted in both adjacent rooms and with normal corridor traffic including student passage, voices, and the noises of locker usage.

For the noisy condition (75–90 decibels) three kinds of noise were used. The first was an external machinery noise created by having a tractor-run power mower pass back and forth outside of the open and curtained windows. The second was a tape recording of *The Blitzkrieg* played in rooms on both sides of the testing room, with the speakers touching the walls. The third kind of noise was human in nature and was created by four male assistants working in the corridors and in adjacent rooms. This noise consisted of running with metal tapped shoes, banging on and slamming wall lockers, talking, whistling, and laughing, banging on walls and blackboards, dragging chairs and desks across the floors, and moving audio-visual equipment through the corridors.

Part 2: Experimental. The experimental section was divided into quiet and noisy conditions. The *Ss* were tested on the stage, which was draped to shield against extraneous noise. Prerecorded white noise issued from a central speaker with *Ss* seated around and equidistant from it. The quiet condition consisted of steady white noise of approximately 50 decibels. The noisy condition consisted of phases of quiet white noise of approximately 50 decibels alternated with phases of loud white noise of approximately 80 decibels. The quiet and noisy periods each consisted of a total of 15 minutes with intervals ranging 30–180 seconds and with a mean interval of 75 seconds.

Tape recordings were made of each entire testing session, and measurements of the actual sound-pressure levels present during the testing sessions were taken every 60 seconds, providing 30 measurements for each condition. Measurements were made with an Industrial Acoustics Company Model SS 375 Sound Spectrometer set on Scale C at fast speed and an H. H. Scott Type 450 Sound Survey Meter set on C weighting. The noise characteristics for each condition are presented in Table 3.

TABLE 1

Testing Conditions by Group

Group	N	Condition: Situation	Noise	Task
1	31	Classroom	Quiet	Homework
2	37	Classroom	Quiet	Test
3	31	Classroom	Average	Test
4	32	Classroom	Average	Homework
5	31	Classroom	Noisy	Test
6	32	Classroom	Noisy	Homework
7	33	Experimental	Quiet	Test
8	36	Experimental	Noisy	Test

TABLE 2

Typical Noise Ranges of the Sample School

Location	Decibel range[a]
Classroom	
Occupied—class silent	54–62
Occupied—normal speech	60–72
Unoccupied—school unoccupied	46–55
Unoccupied—adjacent to band	72–86
Unoccupied—adjacent to chorus	72–78
Unoccupied—classes in adjacent rooms	52–58
Art room—class in session	56–84
Corridor	
During classes	49–58
Between classes	68–89
Study hall (small)	54–65
Cafeteria during a lunch block	76–94

Note.—Readings were taken over 10-minute time intervals and repeated at least once for each range given. A Model SS-375 Sound Spectrometer from the Industrial Acoustics Company set on Scale C (flat from 37.5-9600 cps) on fast speed was used.

[a] Decibels re .0002 microbar.

Supplemental Data

Following testing, the Sarason Test Anxiety Scale for Children (Sarason et al., 1960) and two questionnaires designed to assess *Ss'* perceptions of noise and their awareness of the purpose of the experiment were administered to each *S*. The data from these instruments were used to determine the relationship between individual differences and the effects of noise.

Testing Procedure

The experiment was carried out during the first three periods of two consecutive days to avoid

TABLE 3

NOISE CHARACTERISTICS OF THE ACTUAL EXPERIMENT

Group	Condition	Decibel range	*M*	*SD*
	Classroom			
1	Quiet—Homework	49–58	54.8	1.8
2	Quiet—Test	50–57	54.5	2.2
3	Average—Test	54–72	62.7	5.4
4	Average—Homework	56–72	64.7	4.7
5	Noisy—Test	74–91	82.6	4.7
6	Noisy—Homework	74–90	82.0	4.6
7	Experimental Quiet—Test	50–55	52.6	1.1
8	Experimental Noisy—Test			
	Phase 1—Quiet	52–56	53.2	1.3
	Phase 2—Noisy	79–82	79.9	.9

the contamination of fatigue, which might have increased during the latter part of the day. Quiet and average classroom conditions were run on the first day to avoid feedback of information from *S*s tested to those to be tested. This might have occurred, resulting in a noise set, had the conditions more obviously connected with noise been run first. The *S*s were instructed that they were taking part in a reading project and that the instruments were a means of timing the tests and later checking on the accuracy of timing and instructions.

Treatment of Data

The major body of data was treated in three steps. The means and standard deviations for both speed and accuracy on the reading comprehension test were computed. An *F* maximum test was then used to determine the feasibility of performing analyses of variance. The observed F_{max} numbers were not within the critical regions. Therefore analyses of variances were carried out as the best method of answering the questions raised by the hypotheses.

Data pertaining to the secondary questions were treated descriptively. The secondary questions were those pertaining to individual differences, such as the relationships between pupil perceptions and the effect of noise, or between anxiety and the effects of noise, and were not included in the hypotheses.

RESULTS

Although there was a slight tendency for boys to work faster and to perform less accurately than girls under both of the unfamiliar conditions of white noise, this tendency was of too small a magnitude to be of any practical value. There was no other trend indicating any effect of noise upon performance.

The results of two-way analyses of variance performed on the data for speed and for accuracy indicated that there were no significant differences for condition, sex, or interaction for either speed or accuracy. None of the hypotheses was supported in any degree by the data. Not only were there no significant differences, but there were no trends indicative of any noise effect, detrimental or otherwise.

While there were some minor individual differences, these were not consistent enough or of enough magnitude for practical consideration. The *S*s' perceptions of the effects of noise upon their performance, the degree of noise which was present during the experiment, and the annoyance value of noise had little relationship to actual performance under the noise conditions used. Similarly, measured anxiety had little relationship to actual performance.

DISCUSSION

The major body of data, as treated by analysis of variance, was strong evidence against any effect of noise under the specifications of the experiment and upon the population used. According to existing literature, if a noise effect had been demonstrated, it could have occurred in either the hypothesized detrimental direction or in the opposite direction of assisting performance. Neither effect was demonstrated.

Consideration of the representativeness of the sample, the pertinence of the task to typical school behavior, and the applicability of the conditions used to actual school conditions appeared to warrant generalization outward from the experiment to the public school population in general. At the junior high school level, and possibly at other grade levels, children's tested performance on written tasks, requiring reading comprehension, of the limited duration of a class period in length, is not affected either positively or negatively by the peaks of noise which are typical of a normal school environment.

The effects of noise over time, the effects of noise upon learning, the effects of noise upon tasks of a different nature, and the effects of noise at levels above and be-

low those found in schools were not investigated in the present study.

The major portion of the present study was designed to examine the differences between equated groups of children, rather than the interactions between individual *Ss* and particular noise conditions. The reason for this choice was that practical school planning must consider children in groups. Since the data indicated that school children, as a whole, are not affected detrimentally by noise, it might now be of value to examine individual *Ss* tested under a variety of conditions. While summary inspection of data did not indicate that the results were brought about by extremes of performance cancelling each other out, this does not preclude the possibility that certain individuals might show test-retest changes. An experiment designed to examine this factor, if differences are demonstrated, might provide further insight into possible means of best handling particular children to facilitate the learning process.

REFERENCES

BROADBENT, D. E. Noise, paced performance and vigilance tasks. *British Journal of Psychology*, 1953, **44**, 295–303.

BROADBENT, D. E. Some effects of noise on visual performance. *Quarterly Journal of Experimental Psychology*, 1954, **6**, 1–5.

BROADBENT, D. E. Effects of noise on an "intellectual task." *Journal of the Acoustical Society of America*, 1958, **30**, 824–827. (a)

BROADBENT, D. E. Effects of noise on behavior. In C. M. Harris (Ed.), *Handbook of noise control*. New York: McGraw-Hill, 1958. Pp. 10-1–10-34. (b)

FITZROY, D., & REID, J. L. *Acoustical environment of school buildings: Technical report 1*. New York: Educational Facilities Laboratories, 1963.

GOODENOUGH, F. L. The measurement of mental growth in children. In L. Carmichael (Ed.), *Manual of child psychology*. (2nd ed.) New York: Wiley, 1954. Pp. 459–491.

GRIMALDI, J. V. Sensori-motor performance under varying noise conditions. *Ergonomics*, 1958, **2**, 34–43.

JERISON, H. J. Effects of noise on human performance. *Journal of Applied Psychology*, 1959, **43**, 96–101.

KITAMURA, S. Study of influences of train noise upon schoolchildren: I. *Tohoku Psychologica Folia*, 1964, **23**, 1–2.

LEHMANN, D. W., CRESWELL, W. H., & HUFFMAN, W. J. An investigation of the effects of various noise levels as measured by psychological performance and energy expenditure. *Journal of School Health*, 1965, **35**, 212–214.

PARK, J. F., & PAYNE, M. C., JR. Effects of noise level and difficulty of task in performing division. *Journal of Applied Psychology*, 1963, **47**, 367–368.

PETERSON, A. P. G., & GROSS, E. E., JR. *Handbook of noise measurement*. West Concord, Mass.: General Radio Company, 1963.

SANDERS, A. F. The influence of noise on two discrimination tasks. *Ergonomics*, 1961, **4**, 253–258.

SARASON, S. B., DAVIDSON, K. S., LIGHTHALL, F. F., WAITE, R. R., & RUEBUSH, B. K. *Anxiety in elementary school children*. New York: Wiley, 1960.

SARASON, S. B., & GORDON, E. M. The test anxiety questionnaire: Scoring norms. *Journal of Abnormal Psychology*, 1953, **48**, 447–448.

SUPER, D. E., BRAASCH, W. F., & SHAY, J. B. The effect of distraction on test results. *Journal of Educational Psychology*, 1947, **38**, 373–377.

TEICHNER, W. H., AREES, E., & REILLY, R. Noise and human performance. A psychophysiological approach. *Ergonomics*, 1963, **6**, 83–97.

TERMAN, L. M., & TYLER, L. E. Psychological sex differences. In L. Carmichael (Ed.), *Manual of child psychology*. (2nd ed.) New York: Wiley, 1954. Pp. 1064–1114.

TINKER, M. A. Intelligence in an intelligence test with an auditory distractor. *American Journal of Psychology*, 1925, **36**, 467–468.

WESTON, H. C., & ADAMS, S. The performance of weavers under varying conditions of noise: Industrial Health Research Board, Report Number 70. London: H. M. Stationary Office, 1935.

(Received May 19, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 4, 244–249

EFFECT OF QUESTION LOCATION, PACING, AND MODE UPON RETENTION OF PROSE MATERIAL

LAWRENCE T. FRASE[1]
University of Massachusetts

A factorial design with 128 college *S*s was used to study the effect of question location, question pacing, location of relevant content, and question mode upon the retention of question-relevant and incidental prose material. Retention was highest when questions were placed after paragraphs. Retention increased with the frequency of posttreatment questions, but it decreased with frequent pretreatment questions. It was concluded that frequent postquestioning either shaped or elicited appropriate reading skills while frequent prequestions interfered with prose structure. Frequent questioning, either pre- or posttreatment, yielded precise discrimination between relevant and incidental material which took the form of lowered incidental learning without a corresponding increase in relevant learning. Question mode (multiple-choice or constructed response) had no effect. Relevant material was retained better than the incidental, but incidental retention was relatively high if the incidental material followed the question-relevant material.

Several studies (Frase, 1967; Rothkopf, 1966; Rothkopf & Bisbicos, 1967) have shown that questions improve retention of both relevant and incidental material when they occur *after* the prose paragraphs to which they relate. The reason postquestions work better than prequestions seems to be that they provide cues for the elicitation or shaping of efficient reading behaviors. Postquestions serve more than a review function—they produce nonspecific facilitation of retention over the succeeding material. This study attempted to answer some questions raised by a previous study (Frase, 1967).

Frase (1967), using different materials and *S*s, replicated the results obtained by Rothkopf (1966) showing that the position of the questions and knowledge of results were both important factors in learning from prose. In addition, Frase varied the length of passages between questions and found that even though the total number of questions remained the same, the effect of question pacing tended to be different for retention of relevant material as opposed to incidental material. In accordance with Ausubel's (1963) position concerning meaningful verbal learning, the more the material was broken up by questions the lower the incidental learning. Introducing frequent questions, however, tended to improve retention of the relevant material. The results seemed to confirm both a small step approach, for specific retention, and a large step approach, for general retention. The design of that study precluded determining if the interaction between relevant-incidental retention and pacing of questions was significant. The present study was designed to obtain data on this interaction, and also to determine if there is an interaction between question pacing and the position of questions, that is, whether the skills developed by postquestions are more effectively shaped or maintained as the frequency of questions increases.

In the previous study (Frase, 1967) the relevant material was always located in the second part of each paragraph. If contiguity of question and relevant material is a critical factor for retention, then the occurrence of questions after passages would have been most advantageous for retention of the question-relevant material. On the other hand, retention of relevant material which is located in the first part of the paragraphs should be highest when questions are placed before the paragraphs. The strength and direction of question-content proximity was also explored in the present study.

[1] Now at Bell Telephone Laboratories, Murray Hill, New Jersey.

In the present study the mode of questions (multiple-choice or constructed response) used with the prose passages was also varied. Several authors have noted that the predictability of response class (Rothkopf, 1965), difficulty of a stimulus frame (Faust & Anderson, 1967), or question difficulty (Hershberger & Terry, 1965) is related to retention. On the assumption that a constructed response question is more difficult than a multiple-choice item, it was predicted that retention would be higher when constructed response items were used along with the prose materials.

Although knowledge of results is not given with questions, subjects (*S*s) should be able to answer questions when reading the prose passage, thereby providing their own knowledge of results. Hence, it was predicted that retention of relevant material would be higher than retention of incidental material.

Finally, consistent with the previous studies, it was predicted that questions would improve retention most when they were placed after passages.

METHOD

Subjects

One hundred twenty-eight introductory psychology students participated as a course requirement.

Design

A 2 × 4 × 2 × 2 × 2 factorial design with repeated measures on the last factor was used. The factors were: (*a*) question location—before or after paragraphs, (*b*) question pacing—after every 10, 20, 40, or 50 sentences, (*c*) content location—question-relevant material located in the first or second part of each 10-sentence paragraph, (*d*) question mode—multiple-choice or constructed response, and (*e*) retention items—relevant or incidental to questions used in text.

Factor 2 involved one question after each 10 sentences, two questions after every 20 sentences, four questions after every 40 sentences, or five questions after every 50 sentences. The same questions were thus used under all conditions but the pacing of the questions was changed.

Stimulus Material

A 2000 word passage concerning the life of William James was selected from *Psychology: the Science of Mental Life*, by G. Miller (1962).[2] The passage was divided into 20 paragraphs of 10 lines each. For each 10-line paragraph there were two five-alternative multiple-choice items, one relating to the first half of the paragraph and one relating to the second. These materials were the same as those used in the previous study (Frase, 1967). For one-half of the *S*s the questions relating to the first part of the paragraphs were placed before or after the paragraphs. For the other half of the *S*s, the questions relating to the second part of the paragraphs were placed either before or after the paragraphs. These questions, which *S*s saw when they read the materials, were called *relevant* questions. The other half of the questions, over which *S*s were tested later but which they did not see during reading, were called *incidental.*

The alternative responses were dropped from all the multiple-choice items, yielding sets of relevant and incidental constructed response items. In some cases minor rewording was necessary. For instance, the question:

> There were ____ children in the James family.
> 1. 3
> 2. 4
> 3. 5
> 4. 6
> 5. William was an only child.

became:

> How many children were there in the James family?

The materials were presented in the same manner as in the previous study (Frase, 1967). Each 10-sentence paragraph and each question occurred on a separate sheet of mimeographed 4 × 11 inch paper (a total of 40 pages). The sequence of questions and paragraphs which each *S* saw was determined by the experimental condition to which he had been assigned.

The criterion retention test (40 items) consisted of both the relevant and incidental multiple-choice items already constructed for use with the prose passage. The criterion test was found at the end of the prose materials. The entire package of experimental materials was sealed so that it would not be opened until instructions to do so were given.

Procedure

The experiment was administered to all 128 *S*s in a large auditorium. As *S*s reported for the experiment they were randomly assigned experimental materials and directed to alternate seats. The *S*s were instructed not to look at the materials until told to do so. There were four monitors used to maintain control over *S*s' behavior.

[2] Permission for the experimental use of these copyrighted materials was kindly granted by the publishers, Harper & Row, Inc., 49 East 33rd Street, New York, New York.

After all *S*s had been seated, instructions were read stating that this was an experiment to find out how much people can learn from reading material. The *S*s were told to read each page of material carefully and to turn each page face down after they had read it. They were not to review or look back at any page after they had once read it. They were told to try and answer the questions when they encountered them, and that a final test would be found at the end of the reading material. When they completed the reading task they were to go on to the final test.

The *S*s were asked if they had any questions, after which the instructions were again read. The *S*s were then told to open their materials and to begin.

RESULTS

Question Location

As in the previous studies, questions facilitated retention more when they were placed after the prose passage ($F = 10.43$, $df = 1/96$, $p < .001$). The means were 11.49 and 13.02 for the before and after treatments, respectively.

Retention Items

Retention of the relevant information was significantly higher then retention of the incidental information ($F = 28.6$, $df = 1/96$, $p < .001$). The means for the relevant and incidental information were 12.93 and 11.59, respectively. In the previous study a control group (which merely read through the prose without receiving questions) was run and its mean was equivalent to the average of all experimental groups on the incidental test items. The mean of 11.59 ($N = 128$) is the average of all groups from the present study on incidental test items. This mean is the control mean reported in Figures 1–3 below.

Question Pacing

Figure 1 presents the data on the interaction between question position and pacing of questions (interaction $F = 3.2$, $df = 3/96$, $p < .05$). It can be seen that the advantage of the posttreatment groups became larger the more frequent the questions. Conversely, the disadvantage of placing questions in front of the passages was strongest when the questions occurred most frequently. Posttreatment questions

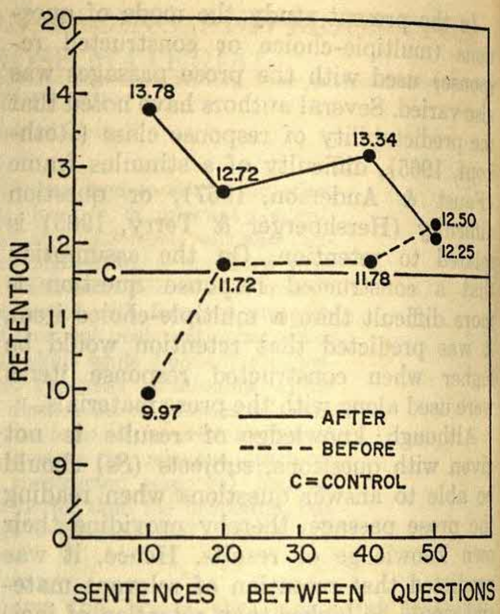


FIG. 1. Retention as a function of question position and pacing.

evidently shape or elicit reading skills, or mathemagenic behaviors (Rothkopf, 1965), more effectively with frequent questions. On the other hand, when placed before passages the questions lose the capacity to arouse and maintain those skills. The extremely low mean for the 10-sentence condition when questions preceded the passages indicates that considerable information was lost. The locus of this lost information is important because, if it were lost from incidental material, it could be stated that prequestions focus attention on relevant material. As a matter of fact, the F ratio for the three-way interaction of Question Location × Pacing × Retention items was fractional. Evidently, frequent prequestions interfered in a similar manner with retention of both relevant and incidental information. The most reasonable conclusion seems to be that frequent prequestioning tends to destroy continuity of the prose materials.

Figure 2 reports the data concerning the interaction of pacing with retention items (interaction $F = 3.24$, $df = 3/96$, $p < .05$). Obviously, retention of the incidental material was depressed with frequent questioning. This depression is statistically

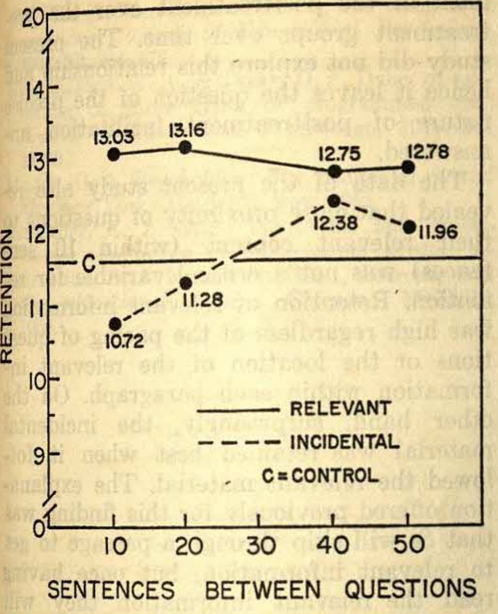


Fig. 2. Retention as a function of retention item and question pacing.

independent of question position and therefore must be due to the size of the passages between questions. The effect of questions evidently becomes more precise, excluding more irrelevant information, as questions become more frequent. The general conclusion seems to be that if questions are frequent enough they will selectively reinforce retention of prose content, whether they come before or after paragraphs. With frequent postquestions general retention is maximized, but at the same time differentiation of relevant from incidental material becomes more precise.

It seems clear that two processes occur when adjunct questions are used effectively with prose materials. First, selective reinforcement of the relevant material, and second, the development of effective reading behaviors. The first process, which depends upon question pacing, is independent of question location; the second process is contingent upon postquestioning.

Content Location

Figure 3 displays the interaction between location of the material and retention item (interaction $F = 5.6$, $df = 1/96$, $p < .025$). If close proximity of question and related material were a critical variable, then there should have been a significant interaction between question location and content location, which there was not. On the contrary, Figure 3 says that, regardless of pacing or location of questions, higher incidental retention was achieved if the incidental material followed the relevant material. This finding suggests that in the previous study (Frase, 1967) part of the observed difference between relevant and incidental retention was due to the design of the materials—the relevant material was always in the second part of each paragraph. To explain the depressed scores when incidental material is located early in the paragraphs, regardless of question location, it is necessary to assume that *S*s know that the relevant material will be located in the later portion of the paragraphs. If this assumption is true, then it can be stated that the results show that *S*s will skip over content (rejecting information) to get relevant information, but once they have gotten relevant information they will not skip over the remaining material.

The assumption that *S*s know the location of relevant and incidental material,

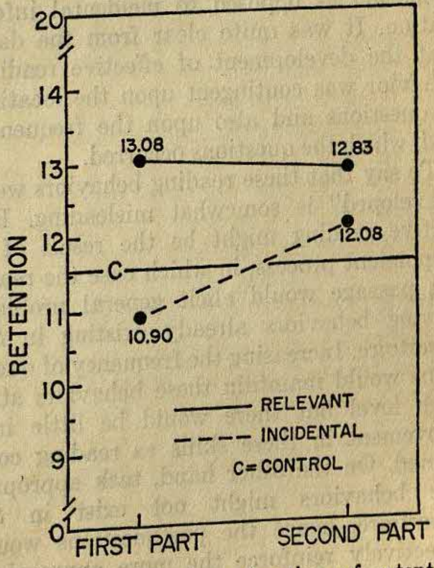


Fig. 3. Retention as a function of retention item and position of content within each 10-sentence paragraph.

whether questions precede or follow paragraphs, seems reasonable since, for any one of the 32 experimental groups, the relevant material was always located in the same position within the paragraphs. Hence, *S*s could learn to expect the relevant material in a certain portion of the passage. The only problem with this assumption, upon which the information rejection explanation rests, is that one might expect an interaction of Content Location × Retention Item × Pacing, that is, questions would have to be frequent in order for *S*s in the postquestion group to discover that the relevant material was located in a certain place within paragraphs.

Question Mode

The final factor explored in the present study—question mode—showed no significant differences between multiple-choice and constructed response items.

DISCUSSION

The present study directly replicated earlier research in two respects. Questions which were placed after prose paragraphs had both a specific and general facilitative effect, and questions had a more powerful effect upon the retention of relevant information as opposed to incidental information. It was quite clear from the data that the development of effective reading behavior was contingent upon the location of questions and also upon the frequency with which the questions occurred.

To say that these reading behaviors were "developed" is somewhat misleading. Effective reading might be the result of a respondent process, in which case the reading passage would elicit general problem solving behaviors already existing in *S*'s repertoire. Increasing the frequency of questions would maintain these behaviors at a high level but there would be little improvement in these skills as reading continued. On the other hand, task appropriate behaviors might not exist in *S*'s repertoire, hence the postquestions would selectively reinforce the more appropriate behaviors. This instrumental learning should be reflected in a gradual improvement of the posttreatment over the pretreatment groups over time. The present study did not explore this relationship and hence it leaves the question of the precise nature of posttreatment facilitation unanswered.

The data of the present study also revealed that close proximity of questions to their relevant content (within 10 sentences) was not a critical variable for retention. Retention of relevant information was high regardless of the pacing of questions or the location of the relevant information within each paragraph. On the other hand, surprisingly, the incidental material was retained best when it followed the relevant material. The explanation offered previously for this finding was that *S*s will skip through a passage to get to relevant information, but once having read the relevant information they will continue to read the remaining information. This explanation is consistent with the view that attention involves information rejection (Berlyne, 1965; Schroder, Driver, & Streufert, 1967), and contrary to the drive reduction hypotheses which asserts that, once having read the relevant information, *S*s' uncertainty is reduced and hence the remaining information will be nonreinforcing.

Although it was predicted that the most difficult questions (constructed response) would lead to highest retention, there was no significant effect of question mode upon retention. The five-alternative multiple-choice questions were rather difficult for *S*s, and it is probable that the form of multiple-choice items used in the present study did not provide an adequate test of the hypothesis. True-false alternatives might have been more appropriate.

REFERENCES

AUSUBEL, D. P. *The psychology of meaningful verbal learning.* New York: Grune & Stratton, 1963.

BERLYNE, D. E. *Structure and direction in thinking.* New York: Wiley, 1965.

FAUST, G. W., & ANDERSON, R. C. Effects of incidental material in a programmed Russian vocabulary lesson. *Journal of Educational Psychology,* 1967, **58,** 3–10.

FRASE, L. T. Learning from prose material: Length

of passage, knowledge of results, and position of questions. *Journal of Educational Psychology*, 1967, **58,** 266–272.

HERSHBERGER, W. A., & TERRY, D. F. Delay of self-testing in three types of programmed text. *Journal of Educational Psychology*, 1965, **56,** 22–30.

MILLER, G. A. *Psychology: The science of mental life.* New York: Harper & Row, 1962.

ROTHKOPF, E. Z. Some theoretical and experimental approaches to problems in written instruction. In J. Krumboltz (Ed.), *Learning and the educational process.* Chicago: McNally, 1965.

ROTHKOPF, E. Z. Learning from written instructive material: An exploration of the control of inspection behavior by test-like events. *American Educational Research Journal*, 1966, **3,** 241–249.

ROTHKOPF, E. Z., & BISBICOS, E. E. Selective facilitative effects of interspersed questions on learning from written materials. *Journal of Educational Psychology*, 1967, **58,** 56–61.

SCHRODER, H. M., DRIVER, M. J., & STREUFERT, S. *Human information processing.* New York: Holt, 1967.

(Received May 22, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 4, 250-255

RETROACTIVE FACILITATION IN MEANINGFUL VERBAL LEARNING

DAVID P. AUSUBEL, MARY STAGER, AND A. J. H. GAITE
Ontario Institute for Studies in Education[1]

In order to ascertain whether retroactive interference occurs in meaningful verbal learning and retention, the experimental conditions favoring such interference were maximized by using both unfamiliar and conflicting original and interpolated learning materials. The effects of interpolated learning (Buddhism) and of overlearning of the original material (Zen Buddhism) were tested in a 2 × 2 factorial design, using 156 12th grade pupils. Both independent variables significantly facilitated the retention of the original material (overlearning: $p < .01$; interpolation: $p < .05$). The facilitating influence of interpolated learning was attributed to the rehearsal and clarification of the original material which it presumably induced. The absence of a significant interaction term indicated that prior overlearning did not differentially affect the later facilitating effect of interpolation.

The phenomenon of retroactive interference in verbal learning has been clearly demonstrated in many studies which have used nonmeaningful and unconnected materials, chiefly nonsense syllables. However, there is much doubt as to whether retroactive interference occurs when connected material is meaningfully learned (i.e., when it interacts on a nonarbitrary, substantive basis with established ideas in cognitive structure).

In general, those studies with connected material which demonstrated the occurrence of retroactive interference have demanded verbatim recall of material (e.g., Jenkins & Sparks, 1940; King & Cofer, 1960; Slamecka, 1959, 1960a, 1960b, 1962). Further, Mehler, and Miller (1964) obtained retroactive interference for the syntactic, but not for the semantic, aspects of potentially meaningful sentences, and Newman (1939) demonstrated retroactive interference for nonessential, but not for essential, details of a narrative.

The majority of studies (Ausubel, Robbins, & Blake, 1957; Hall, 1955; McGeoch & McKinney, 1934; Mehler & Miller, 1964; Newman, 1939) requiring substantive (as opposed to verbatim) recall of connected verbal material have failed to demonstrate clearly the operation of retroactive interference. Two of these studies (Ausubel et al., 1957; Mehler & Miller, 1964), in fact, found that material similar to the original material and interpolated between original learning and the tests for retention of such learning led to retroactive facilitation. But a recent study by Entwisle and Huggins (1964) indicated that, when engineering students were tested on a set of principles in electrical circuit theory, the interpolation of a highly similar set of principles before testing produced significant retroactive interference. It is debatable, however, whether the type of learning involved was nonarbitrary and substantive in nature; it is quite possible that the students may have learned the material, which was essentially mathematical (formulae, etc.) rather than verbal, by rote. Hence, it is concluded that there has been no definitive demonstration of the retroactive interference phenomenon in studies requiring the meaningful (nonarbitrary and substantive) learning and retention of connected verbal material.

Traditionally, retroactive interference has been explained in behavioristic terms: Specific responses (from the orginally learned material) are lost (forgotten) because they are superseded by competing associative tendencies (from the interpolated material) having greater relative strength. A major variable in determining the amount of forgetting is the similarity of responses in original and interpolated

[1] The authors are also at the University of Toronto.

activities (Osgood, 1953) where increased similarity, short of identity, leads to increased interference. This relationship has been experimentally verified only in studies of rote learning.

In situations involving the substantive retention of potentially meaningful material, the applicability of the behavioristic explanation is questionable for, as noted earlier, there has been no agreement concerning the effects on retention of similar interpolated materials. In one such situation, Ausubel et al. (1957) found that an interpolated passage, which compared Buddhism and Christianity and which was substantively similar to the Buddhism passage that was tested, induced retroactive facilitation. This result suggested the interpretation that in the case of meaningful learning, where new concepts and propositions are nonarbitrarily and substantively related (anchored) to existing ideas in cognitive structure, the newly learned material is protected, by virtue of such anchorage, from the interfering effects of subsequently encountered competing stimuli and responses (Ausubel et al., 1957). More important than the similarity variable for learning and retention in these circumstances, it was hypothesized, are such variables as the availability of relevant anchoring ideas in cognitive structure, their stability and clarity, and their discriminability from the learning material.

These authors proposed therefore that the influence of interpolated learning on retention is not necessarily a function of similarity of original and interpolated materials; instead it depends on whether or not the interpolated passage increases or decreases the discriminability of the original passage from its anchoring concepts in cognitive structure and hence counteracts or promotes irreversible reduction (forgetting), that is, the process whereby the originally learned material is reduced to a least common denominator of and thus is no longer dissociable (retrievable) from the ideational system in which it is embedded.

The present study was designed to discover whether retroactive interference could be demonstrated in a learning situation that involved the substantive retention of potentially meaningful material and that was more analogous both to the Entwisle and Huggins (1964) study and to experiments demonstrating retroactive interference with verbatim recall than was the Ausubel et al. (1957) study. To satisfy these conditions, the original and interpolated materials had to be both unfamiliar and sufficiently similar to each other to engender confusion and conflict. Thus, the interpolated passage, a discussion of Buddhism, was highly similar to and conflicted with basic concepts in the originally learned passage (which dealt with Zen Buddhism), and both passages were generally unfamiliar to the experimental sample.

The present experiment also investigated the operation of another variable whose effect on retroactive interference with the retention of nonmeaningful material and with the verbatim recall of connected material (Slamecka, 1959, 1960a) is well established. It is generally accepted (Slamecka & Ceraso, 1960) that susceptibility to retroactive interference after rote learning is inversely related to the level of verbatim acquisition of the original material. To investigate the effect of overlearning on retroactive interference or facilitation in a meaningful learning context, this experiment was designed so that certain *Ss* reread the original Zen Buddhism passage.

METHOD

Subjects

The experimental sample consisted of 156 students (91 male and 65 female). The *Ss* were drawn from the total membership of all of the Grade 13 classes in two suburban high schools and consisted of those students who were present for all four sessions of the experiment. These sessions, given once at each school on the same day, took a maximum of 40 minutes and were conducted during regular school hours over a period of 1 week.

Learning Passages and Measuring Instrument

The material used to investigate retroactive interference consisted of two passages that on the basis of content analysis were thought to be highly

similar and conflicting. The first (original) passage (approximately 2,200 words in length) was concerned with the history, sacred literature, doctrine, and ethical teachings of Zen Buddhism. The second (interpolated) passage (approximately 2,100 words in length) dealt with similar topics in the Buddhist faith.

A third passage (approximately 1,500 words in length), which dealt with the causes and types of drug addiction, was presented instead of the Buddhism passage to control group *Ss*. Because of its totally different content, it was presumed that this passage would not interfere with the Zen Buddhism passage.

The material in all three passages was selected on the basis of its unfamiliarity to almost any high school student. Hence the interpolated material (the Buddhism and drug addiction passages) differed for the experimental and control groups only in degree of similarity to the original (Zen Buddhism) passage and not, presumably, in familiarity. Empirical confirmation of the unfamiliarity of these passages was obtained when naïve *Ss* who had not studied the material in question made scores on the respective tests that were not significantly better than chance.

A 35-item multiple-choice test on Zen Buddhism was used to measure the learning performance of all *Ss*. Before the data were analyzed, it was decided to eliminate four items from the test. These were items on which the experimental sample did more poorly than chance. In addition, two of these items had negative indexes of discrimination, that is, *Ss* in the bottom quartile (as determined from total test scores) performed better on these two items than did *Ss* in the top quartile. The corrected split-half reliability of this shortened (31-item) version of the test was .73. Scores showed a satisfactory range of variability and their distribution did not deviate significantly from the normal curve.

Procedure

At the beginning of each 40-minute session, *Ss* spent approximately 5 minutes reading instructions. The balance of the session was available for reading the passages and, in the final session, for taking the test of Zen Buddhism; no *Ss* appeared to have difficulty in completing either type of task.

In the first session of the experiment, all *Ss* studied the Zen Buddhism passage. They were told, with this passage and with the Buddhism and drug addiction passages, that they were to read at their customary speed, that they were not to turn back once they had completed reading a page, and that they would be examined on the material at a later time by means of a multiple-choice test. (They were not actually tested on any material other than the Zen Buddhism passage, but the anticipation of a test on each passage was thought necessary to sustain and equate motivation in all conditions.)

After the first session, *Ss* were assigned to one of four groups, according to a 2 × 2 factorial design. Groups A and D were to receive the overlearning treatment, whereby they restudied the Zen Buddhism passage during the second session. Groups A and B were to receive the interpolated learning treatment, studying the Buddhism material in the third session.

TABLE 1

Mean Scores, Cell Variances, and Frequencies of Four Treatment Groups on Test of Zen Buddhism

Group	IBP	OZBP	M	S^2	N
A	present	present	16.08	19.41	38
B	present	absent	11.36	14.31	39
C	absent	absent	9.89	16.46	37
D	absent	present	14.55	16.61	42

Note—Abbreviated: IBP = Interpolated Buddhism passage, OZBP = Overlearning of Zen Buddhism passage.

Two stipulations were made in assigning *Ss* to treatment groups. First, because high school girls have been found to have higher verbal ability than high school boys (e.g., superior performance on the verbal portion of the School and College Ability Test, as shown by Ausubel & Fitzgerald, 1962), *Ss* were assigned to groups in such a way that the ratio of girls to boys in each group was equal. (A chi-square test showed that, after eliminating those *Ss* who were not present for all four sessions, the male-female ratio in each treatment group did not depart significantly from equality.) Second, because of possible differences in ability between the populations of the two schools, equal proportions of students from each school were assigned to each treatment group. (A chi-square test showed that, with the 156 *Ss* present during the entire experiment, the proportion of *Ss* from each school in each group was not significantly different.) Aside from these two restrictions, assignment of *Ss* to groups was made on a random basis.

The second session took place two days after the first. Groups A and D studied the Zen Buddhism passage a second time; Groups B and C studied the unrelated drug addiction passage.

In the third session, 2 days later, Groups A and B studied the potentially interfering Buddhism passage, and Groups C and D studied the unrelated drug addiction passage.

During the final session, which took place 3 days after the third, and 1 week after the initial session, all *Ss* were tested on the Zen Buddhism passage. They were instructed to answer all questions, and not to turn back once a page was completed.

Results and Discussion

Effect of Interpolated Learning on Retention

A comparison of the means (Table 1) on the 31-item test of Zen Buddhism indi-

cates that the interpolated Buddhism passage, when compared with the irrelevant drug addiction passage, did not interfere with, but in fact facilitated, retention of the original Zen Buddhism passage.

Analysis of variance, following Winer's (1962) method for dealing with unequal cell frequencies when cell variances are homogeneous, shows that the overall facilitating effect of the similar interpolated passage is significant, $F = 5.10$, $df = 1/52$, $p < .05$. A nonsignificant interaction term indicates that the interpolated Buddhism material affected retention in the same way regardless of degree of original learning.

The evidence is clear, therefore, that in this meaningful learning situation, retroactive interference did not occur when a connected and potentially meaningful passage was interpolated between material to which it was highly similar and a test for substantive retention of such material. Indeed, the interpolated passage appears to have had an effect that was small but reliably facilitating in comparison to that produced by a dissimilar and nonconflicting alternative passage. Thus it is suggested that the learning of the Buddhism passage may have served as a review and clarification of the Zen Buddhism material, thereby increasing both its stability and clarity and its discriminability from its anchoring concepts (presumably, related aspects of Judaism and Christianity) in cognitive structure.

The interpolation of similar and conflicting material between the meaningful learning of a Zen Buddhism passage and a later test of its retention may have had a facilitating effect on the retention of the Zen Buddhism material because it induced *S*s to compare, on their own, the two sets of material, and thus (*a*) to delineate those similarities between them that define their common differences from those established ideas in cognitive structure to which both were related in the course of learning, and (*b*) to delineate the differences between them. Both of these comparative operations conceivably could have facilitated retention of the original learning material by clarifying and sharpening its distinctive features and by increasing its discriminability from anchoring ideas in cognitive structure. In addition, these comparative operations may have further enhanced retention of the Zen Buddhism material because they necessarily required rehearsal of this material, which rehearsal, in turn, increased its stability, clarity, and discriminability.

To the extent that both the interpolated material and the original learning material share certain common ideas that are differentiable on the same basis from their common anchoring concepts in cognitive structure, later exposure to the interpolated material may increase the possibility that basic differences between the original material and the anchoring ideas will be cognized. In other words, if learning passages B and C are conflicting (similar but not identical), and thus necessarily share certain common differences relative to their common anchoring concepts (A), exposure to passage C makes possible the delineation of a common set of differences between the learning passages (B and C) and A, and may thereby make B more discriminable from A than if later exposure to C had not taken place. On the other hand, comparative efforts aimed at delineating the more specific kinds of differences between original and interpolated materials presumably help to sharpen the distinctive features of the original material, and may thus indirectly increase its discriminability from anchoring ideas in cognitive structure.

Furthermore, in the process of identifying and clarifying simple differences between original and interpolated learning passages, as well as more generic differences which differentiate both learning passages from relevant anchoring ideas in cognitive structure, *S* must necessarily rehearse (i.e., activate or attempt to retrieve from storage) the original learning material. Such rehearsal may facilitate later retention in two ways. First, the very activation or retrieval of partially forgotten material (material in the process of undergoing obliterative reduction) serves as a partial review of this material, and may thus enhance its stability and clarity in cognitive structure (thereby directly in-

creasing its availability or retrievability at the time of later testing). Second, the greater clarity and stability of the original material resulting from rehearsal may indirectly increase its later retrievability by enhancing its discriminability from those established ideational systems in cognitive structure to which it is anchored.

These findings, if replicated and given greater generality, would have far-reaching implications for classroom teaching practice. Instead of suggesting (as do the classical retroactive interference findings in the case of rote learning and retention) that teachers scrupulously avoid introducing similar and conflicting material after a typical lesson involving meaningful learning, they imply that such material should be introduced deliberately. This recommendation would be based on the expectation that conflicting interpolated material would encourage the learner to compare related ideas in the original and interpolated sets of material, and thus facilitate retention of the original material through the influence of such intervening variables as rehearsal and clarification.

Effect of Overlearning on Retention

It is evident from Table 1 that a second session of studying the Zen Buddhism passage improved the retention of this passage relative to the groups who studied it only once. Analysis of variance indicates that this facilitating effect is significant, $F = 49.90$, $df = 1/152$, $p < .01$. This finding is consistent with the results of previous studies (Ausubel & Youssef, 1965; Reynolds & Glaser, 1964) which demonstrated that review facilitates the retention of meaningfully learned material. It presumably does so through mechanisms similar to, but by no means identical with, those postulated above to account for the effects of rehearsal. For one thing, since overlearning of the original material involves both more complete and more explicit repetition than that involved in rehearsal, the facilitating effect of overlearning is accordingly much more pronounced (see Table 1).

The absence of any interaction effect, however, indicates that the interpolation of conflicting material in this experiment did not differentially affect retention of the original material for the groups which overlearned the latter material and the groups which did not. That is, the facilitating effect of interpolation was neither greater nor less when it was preceded by overlearning of the original material than when it was not so preceded, and hence cannot be attributed in any way to the effect of such prior overlearning.

Although this finding contrasts markedly with the comparable, previously discussed situation in regard to rote learning, where overlearning of the original material has been invariably found to diminish subsequent susceptibility to retroactive interference, it is nonetheless readily understandable. When, as in the case of rote learning, interpolation has an interfering effect on the retention of the original material, any factor (e.g., overlearning) that increases the associative strength of such material quite naturally tends to lessen the interfering potential of competing associative tendencies. But since the interpolation of conflicting material facilitates rather than interferes with the retention of meaningfully learned original material, one cannot expect overlearning of the latter material to interact with the very different effects of interpolation in the same way as in the case of rote learning. However, the fact that the facilitating effect of interpolation is as great in a context of prior overlearning as in the absence of such a context, permits the inference that the mechanisms underlying the facilitating influence of interpolation are different than those underlying the facilitating influence of overlearning, and hence that the occurrence of the prior facilitating effect of overlearning does not preclude the later facilitating effect of interpolation.

REFERENCES

Ausubel, D. P., & Fitzgerald, D. Organizer, general background, and antecedent learning varia-

bles in sequential verbal learning. *Journal of Educational Psychology*, 1962, **53**, 243–249.

AUSUBEL, D. P., ROBBINS, L. C., & BLAKE, E. Retroactive inhibition and facilitation in the learning of school materials. *Journal of Educational Psychology*, 1957, **48**, 334–343.

AUSUBEL, D. P., & YOUSSEF, M. The effect of spaced repetition on meaningful learning. *Journal of General Psychology*, 1965, **73**, 147–150.

ENTWISLE, D. R., & HUGGINS, W. H. Interference in meaningful learning. *Journal of Educational Psychology*, 1964, **55**, 75–78.

HALL, J. F. Retroactive inhibition in meaningful material. *Journal of Educational Psychology*, 1955, **46**, 47–52.

JENKINS, J. G., & SPARKS, W. M. Retroactive inhibition in foreign language study. *Psychological Bulletin*, 1940, **37**, 470. (abstract)

KING, D. J., & COFER, C. N. Retroactive interference in meaningful material as a function of the degree of contextual constraint in the original and interpolated learning. *Journal of General Psychology*, 1960, **63**, 145–158.

MCGEOCH, J. A., & MCKINNEY, F. The susceptibility of prose to retroactive inhibition. *American Journal of Psychology*, 1934, **46**, 429–436.

MEHLER, J., & MILLER, G. A. Retroactive interference in the recall of simple sentences. *British Journal of Psychology*, 1964, **55**, 295–301.

NEWMAN, E. B. Forgetting of meaningful material during sleep and waking. *American Journal of Psychology*, 1939, **52**, 65–71.

OSGOOD, C. E. *Method and theory in experimental psychology*. New York: Oxford University Press, 1953.

REYNOLDS, J. H., & GLASER, R. Effects of repetition and spaced review upon retention of a complex learning task. *Journal of Educational Psychology*, 1964, **55**, 297–308.

SLAMECKA, N. J. Studies on retention of connected discourse. *American Journal of Psychology*, 1959, **72**, 409–416.

SLAMECKA, N. J. Retroactive inhibition of connected discourse as a function of practice level. *Journal of Experimental Psychology*, 1960, **59**, 104–108. (a)

SLAMECKA, N. J. Retroactive inhibition of connected discourse as a function of similarity of topic. *Journal of Experimental Psychology*, 1960, **60**, 245–249. (b)

SLAMECKA, N. J. Retention of connected discourse as a function of duration of interpolated learning. *Journal of Experimental Psychology*, 1962, **63**, 480–486.

SLAMECKA, N. J., & CERASO, J. Retroactive and proactive inhibition of verbal learning. *Psychological Bulletin*, 1960, **57**, 449–475.

WINER, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

(Received May 30, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 4, 256-260

DIFFERENTIAL PREDICTION OF ACADEMIC ACHIEVEMENT IN CONFORMING AND INDEPENDENT SETTINGS[1]

GEORGE DOMINO
Fordham University

The grade-point average (GPA) of 4 groups of college juniors, representing high and low scorers on the CPI Ac and Ai scales, was analyzed to test the hypothesis that conforming and independent achievement motivation (as measured by the CPI) is related to scholastic achievement reflective of conforming or independent behavior. Specific hypotheses regarding differential achievement as a function of Ac and Ai scores were tested and, in general, supported. The results obtained underscore the heuristic value of separating GPA into subcategories reflective of diverse demands.

Previous studies with the California Psychological Inventory (CPI) have clearly demonstrated its usefulness in predicting academic achievement in various educational settings and with differing samples. CPI scales have shown considerable validity in predicting performance in mathematics (Keimowitz & Ansbacher, 1960), in introductory psychology (Gough, 1964b), in medical school (Gough & Hall, 1964), in high school (Gough, 1964a; Snider, 1966), with gifted pupils (Lessinger & Martinson, 1961), students of average ability (Fink, 1962; Gough & Fink, 1964), military enlisted personnel (Rosenberg, McHenry, Rosenberg, & Nichols, 1962), National Merit finalists (Holland, 1959), and other samples.

Although all of the above studies reported positive findings, the differential predictive potential of the CPI was not maximized since the underlying nature of the criterion (grade-point average—GPA, or honor-point ratio) was disregarded. As any student can well document, equivalent grades in different courses do not represent equivalent performances. A student's academic performance reflects a variety of factors, including personality aspects that can enhance or interfere with optimal functioning in settings where conformity or independence are differentially rewarded.

Gough (1957), in constructing the CPI, has explicitly recognized this by including two scales of achievement motivation. The first of these, Achievement via Conformance (Ac), identifies those aspects of motivation that facilitate achievement in settings where conforming behavior such as acceptance of regulations, a high degree of self-discipline, efficiency, and responsibility are rewarded. The second scale, Achievement via Independence (Ai), identifies those motivational aspects that facilitate achievement in settings rewarding independence, individuality, self-reliance, and creative innovation.

The present study is an attempt to relate these personality measures of conforming and independent motivation to scholastic achievement attained in a setting rewarding conforming behavior, and in a setting rewarding independent behavior, to test the hypothesis that the Ac and Ai scales show differential predictive patterns in different settings.

METHOD

A sample of 348 liberal arts juniors attending a California state college on a full-time basis were administered a test battery, including the Ac and Ai scales of the CPI, and the D 48, a nonverbal test of intelligence (Domino, 1964; Gough & Domino, 1963). The distributions of scores on the Ac and Ai scales were tallied in order to select four groups (*a*) students scoring high on both scales (HiAc-HiAi); (*b*) students scoring high on Ac but low on Ai (HiAc-LoAi); (*c*) students scoring low on Ac but high on Ai (LoAc-HiAi); and (*d*) students scoring low on both scales (LoAc-LoAi).

[1] The author is indebted to Harrison Gough for suggestions in regard to various aspects of this paper, to John Walsh for statistical assistance, and to the Fordham University Computer Center for granting computer time.

Registrar's records were then consulted to determine courses taken and grades received by these *Ss* during their first 2 years of college. Only *Ss* enrolled in full programs (15 units or more per semester) for four consecutive semesters were considered. For every course taken by any student the instructor was interviewed in an attempt to determine whether the particular course rewarded conforming or independent behavior on the part of the students.

A course was deemed as rewarding conforming behavior if it was characterized by emphasis on: (*a*) memorizing of technical terms, definitions, poems, etc.; (*b*) presentation of material through lectures; (*c*) objective type examinations; (*d*) keeping of attendance records; (*e*) discipline and adherence to regulations (e.g., no smoking, absences justified by written medical reasons); (*f*) clearly defined and frequent homework assignments emphasizing convergent thinking; (*g*) rare use of visual aids, outside speakers, little variation in class routine; (*h*) close correspondence between lecture material and textbook; (*i*) identical assigned readings for all class members; and (*j*) course grade determined by proportional weighting of various course requirements.

A course was deemed as rewarding independent behavior if it was characterized by emphasis on: (*a*) ideas rather than facts; (*b*) seminar discussions, student presentations, or question-answer format; (*c*) no examinations, or examinations involving essay questions; (*d*) little concern for attendance; (*e*) little explicit emphasis on discipline and adherence to school regulations; (*f*) no homework assignments, or assignments demanding divergent thinking; (*g*) variety of presentation, as indicated by use of visual aids, tape recordings, outside speakers, or other material; (*h*) little direct overlap between class discussions and textbook content; (*i*) suggested readings, or assigned readings individually tailored to a student's interests; and (*j*) grade determined by consultation with student or by global evaluation of student's performance.

Using these criteria, it was possible to label 73 courses as conforming and 32 as independent.

Every student's grades were divided into those received in conforming courses and those received in independent courses.[2] Since it was not possible to contact all instructors concerned, and since some students had taken only one type of course (typically, conforming courses), a number of students had to be omitted from the analysis.

Four groups of 22 *Ss* each were finally retained; the groups were matched for sex and intelligence (D 48 scores). Table 1 indicates the Ac and Ai scores descriptive of each group, as well as the sex ratios and D 48 scores.

[2] All grades were converted by an honor-point ratio formula where A = 4, B = 3, C = 2, D = 1, F = 0; grades were multiplied by credits per course and divided by total credits carried.

TABLE 1

SUMMARY STATISTICS FOR FOUR ACHIEVEMENT GROUPS

Variables	Achievement groups			
	HiAc-HiAi	HiAc-LoAi	LoAc-HiAi	LoAc-LoAi
CPI Ac range	30–35	30–35	16–22	16–22
$\bar{X}$	31.9	31.5	19.5	19.1
SD	1.89	1.50	2.28	2.10
CPI Ai range	23–28	11–16	23–25	11–16
$\bar{X}$	24.5	14.6	23.9	14.2
SD	1.67	1.42	0.78	1.37
Sex composition	19 M	17 M	18 M	19 M
	3 F	5 F	4 F	3 F
D 48				
$\bar{X}$	26.8	27.2	27.0	25.9
SD	5.7	6.0	5.8	6.2

Note.—N = 22 per group.

The following specific hypotheses were made:

1. Concerning total GPA (GPAt):
   a. the HiAc-HiAi group should have a higher mean GPAt than any of the other groups;
   b. the LoAc-LoAi group should have a lower mean GPAt than any of the other groups.
2. Concerning conforming GPA (GPAc):
   a. the HiAc-HiAi group should have a higher mean GPAc than the LoAc-HiAi group;
   b. the HiAc-LoAi group should have a higher mean GPAc than the LoAc-LoAi group.
3. Concerning independent GPA (GPAi):
   a. the HiAc-HiAi group should have a higher mean GPAi than the HiAc-LoAi group;
   b. the LoAc-HiAi group should have a higher mean GPAi than the LoAc-LoAi group.

These hypotheses were tested by means of F ratios across the four mean differences for each of the three GPAs, as shown in Table 2. Specific intergroup comparisons (t tests) were then carried out to evaluate the indicated comparisons.

RESULTS

Table 2 presents the $\bar{X}$s, *SD*s, and F ratios for the intergroup comparisons of GPAt, GPAc, and GPAi.

All four F ratios achieved statistical significance at the .01 level; t tests for individual comparisons are therefore permissible. Of the nine t test comparisons in Table 3, involving the six specified hypotheses, seven reached statistical significance and two were in the hypothesized direction although not significant.

For overall GPA, the HiAc-HiAi subsample was significantly higher than any

TABLE 2
GROUP COMPARISONS ON THREE TYPES OF GRADE-POINT AVERAGE (GPA)

Group	Conforming GPA		Independent GPA		Total GPA	
	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*
HiAc-HiAi	2.97	.55	3.33	.37	3.15	.38
HiAc-LoAi	2.66	.93	2.35	.77	2.50	.78
LoAc-HiAi	2.49	.45	2.70	.59	2.60	.48
LoAc-LoAi	2.34	.48	2.14	.51	2.24	.44
F test	6.77*		16.40*		9.98*	

* $p < .01$.

other. For conforming GPA, the HiAc-HiAi subsample was significantly higher than the LoAc-HiAi. For independent GPA, the HiAc-HiAi sample was higher than the HiAc-LoAi, and the LoAc-HiAi was higher than the LoAc-LoAi. The LoAc-LoAi sample was significantly lower on total GPA than any other group.

DISCUSSION

In view of repeated findings (cf. Gough, 1966) that the roles of Ac and Ai as forecasters of scholastic achievement vary according to intellectual ability and environmental demands, it is important to note that *S*s in this study represent average college juniors, as indicated by a mean GPAt of 2.62 across all four groups, as well as the D 48 distribution which approximates the average for college students (Domino, 1964).

It should also be acknowledged that the restrictions imposed in forming the four groups negate any possibility of randomness of sampling. In addition, use of junior-year *S*s automatically restricted the range of possible grades obtained in the first two college years, since failing and/or marginal students would have been eliminated.

Liberal arts juniors were specifically selected since their scholastic index was based on four semesters of work which would include general courses, exposure to both sciences and humanities, and a predominance of nonmajor subjects.

It is clear that some disciplines are more amenable to one type of presentation than another. In fact, one may question whether the obtained results reflect subject matter

TABLE 3
INTERGROUP COMPARISONS

Grade-Point Average	All courses			Humanities			Sciences		
	$\bar{X}$	*SD*	*t*	$\bar{X}$	*SD*	*t*	$\bar{X}$	*SD*	*t*
Total									
HiAc-HiAi vs.	3.15	.38		3.40	.52		2.90	.20	
HiAc-LoAi	2.50	.78	3.34**	2.46	.69	5.10**	2.54	.73	2.23*
LoAc-HiAi	2.60	.48	4.06**	2.73	.36	4.96**	2.47	.38	4.69**
LoAc-LoAi	2.24	.44	7.03**	2.20	.41	8.51**	2.28	.45	5.90**
LoAc-LoAi vs.	2.24	.44		2.20	.41		2.28	.45	
HiAc-LoAi	2.50	.78	1.32	2.46	.69	1.48	2.54	.73	1.42
LoAc-HiAi	2.60	.48	2.45**	2.73	.36	4.57**	2.47	.38	1.52
Conforming									
HiAc-HiAi vs.	2.97	.55		2.88	.54		3.06	.68	
LoAc-HiAi	2.49	.45	3.04**	2.51	.38	2.62**	2.47	.61	3.02**
HiAc-LoAi vs.	2.66	.93		2.60	.85		2.72	.76	
LoAc-LoAi	2.34	.48	1.37	2.66	.30	0.31	2.02	.52	3.57**
Independent									
HiAc-HiAi vs.	3.33	.37		3.68	.46		2.98	.32	
HiAc-LoAi	2.35	.77	5.15**	2.30	.61	8.46**	2.40	.58	4.11**
LoAc-HiAi vs.	2.70	.59		2.78	.46		2.62	.59	
LoAc-LoAi	2.14	.51	3.26**	2.26	.38	4.09**	2.02	.49	3.68**

* $p < .05$.
** $p < .01$.

differences rather than the interaction of style of achievement with instructor's style of teaching. The answer to this question was obtained in two ways: (*a*) Catalog descriptions of the 105 courses were independently rated by three judges, not acquainted with this study, as either humanities or science courses. Of the 83 courses labeled as humanities by at least two of the three judges, 26 had been designated as independent and 57 as conforming. Of the 22 science courses, 6 had been designated as independent and 16 as conforming. A chi-square analysis, incorporating Yates' correction, gave a nonsignificant value of .003. (*b*) The same intergroup comparisons made on GPAt, GPAc, and GPAi (see Table 3) were computed for humanities courses only and for science courses only. The results of these comparisons, also presented in Table 3, are essentially identical, with the exception of the LoAc-LoAi vs. LoAc-HiAi comparison which for science courses does not reach statistical significance.

The results obtained underscore the heuristic value of separating GPA into subcategories reflective of diverse demands. In this connection, it should be noted that Elton (1966) analyzed Omnibus Personality Inventory scores for 92 freshmen students having identical departmental schedules, and reported a slight tendency for personality traits predictive of grades in English courses to be negatively related to grades in chemistry.

Given current curricular and attitudinal changes occurring on most college campuses, it is important to keep in mind that not every student can achieve his best in a conformist (or independence-demanding) setting. In the past, the typical curriculum may have demanded and rewarded conforming behavior. Today, wider use of honors programs, undergraduate seminars, interdepartmental majors, 3-year baccalaureate programs, and other curricular reforms, seem to emphasize independent behavior. Rather than fit the student to the curriculum as is presently done, it might be extremely worthwhile to fit the curriculum to the student by providing each student with the type of setting which most effectively utilizes his potential.

In commenting on the results obtained in this study, it must be noted that they derive from a particular college setting and may not be generalizable to other educational institutions. Both Holland (1959) and Thistlethwaite (1959) have presented evidence of variation in institutional environments, leading to somewhat different patterning of variables predictive of academic achievement within these settings. The HiAc-HiAi student, very likely, will do well in any academic environment. For the HiAc-LoAi and LoAc-HiAi students, however, there is a distinct and understandable interaction between achievement and the demands of the environment.

REFERENCES

Domino, G. Comparison of the D 48, Cattell Culture Fair, and Army Beta tests in a sample of college males. *Journal of Consulting Psychology*, 1964, **28**, 468–469.

Elton, C. F. The relationship of grades to personality. *Psychology in the Schools*, 1966, **3**, 180–183.

Fink, M. B. Objectification of data used in underachievement self-concept study. *California Journal of Educational Research*, 1962, **13**, 105–112.

Gough, H. G. *Manual for the California Psychological Inventory*. Palo Alto, Calif.: Consulting Psychologists Press, 1957.

Gough, H. G. Academic achievement in high school as predicted from the California Psychological Inventory. *Journal of Educational Psychology*, 1964, **55**, 174–180. (a)

Gough, H. G. Achievement in the first course in psychology as predicted from the California Psychological Inventory. *Journal of Psychology*, 1964, **57**, 419–430. (b)

Gough, H. G. Graduation from high school as predicted from the California Psychological Inventory. *Psychology in the Schools*, 1966, **3**, 208–216.

Gough, H. G., & Domino, G. The D 48 test as a measure of general ability among grade school children. *Journal of Consulting Psychology*, 1963, **27**, 344–349.

Gough, H. G., & Fink, M. B. Scholastic achievement among students of average ability, as predicted from the California Psychological Inventory. *Psychology in the Schools*, 1964, **1**, 375–380.

Gough, H. G., & Hall, W. B. Prediction of performance in medical school from the California Psychological Inventory. *Journal of Applied Psychology*, 1964, **48**, 218–226.

Holland, J. L. The prediction of college grades from the California Psychological Inventory and

the Scholastic Aptitude Test. *Journal of Educational Psychology*, 1959, **50**, 135–142.

Koimowitz, R. I., & Ansbacher, H. L. Personality and achievement in mathematics. *Journal of Individual Psychology*, 1960, **16**, 84–87.

Lessinger, L. M., & Martinson, R. A. The use of the California Psychological Inventory with gifted pupils. *Personnel Guidance Journal*, 1961, **39**, 572–575.

Rosenberg, L. A., McHenry, T. B., Rosenberg, A. M., & Nichols, R. C. The prediction of academic achievement with the California Psychological Inventory. *Journal of Applied Psychology*, 1962, **46**, 385–388.

Snider, J. G. Academic achievement and underachievement in a Canadian high school as predicted from the California Psychological Inventory. *Psychology in the Schools*, 1966, **3**, 370–372.

Thistlethwaite, D. College environments and the development of talent. *Science*, 1959, **130**, 71–76.

(Received June 12, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 4, 261-266

EFFECTS OF CHALLENGING AND SUPPORTIVE INSTRUCTIONS ON VERBAL LEARNING IN OLDER PERSONS

EDITH ROSS

Teachers College, Columbia University

60 young and 60 old Ss were given two paired-associate learning tasks differing in level of difficulty under neutral, supportive, and challenging instructions. Older Ss revealed a greater performance decrement on the more difficult task than did the younger group. Elderly Ss did least well on the acquisition phase of learning under challenging instructions and best under the supportive treatment. No differential effect of instructions on relearning was found. The difference between young and old in number of trials needed to master the material under the supportive condition was significantly smaller than under the challenging one. The effects of feelings of inadequacy aroused by the ego-involving instructions on the difficult task were discussed in relation to their interference with the performance of the elderly Ss.

Research evidence, accumulated over the past three decades, appears to support the hypothesis that elderly *S*s, in comparison with younger ones, exhibit performance deficits on a wide variety of learning tasks. However, the magnitude of the decrement is believed to be dependent upon the nature of the material to be learned. Studies by Gilbert (1941), and Korchin and Basowitz (1957) have indicated that the elderly seem to be less capable than the young of dealing with novel and difficult material which cannot be readily integrated with earlier experience. "Old" learning and retention is believed to be less subject to deterioration. Testing memory function on a number of different variables, Gilbert (1941) discovered that the greatest decrement in the performance of older persons appeared in the learning of a Turkish-English vocabulary and in the acquisition and retention of paired-associates. The difficulty encountered in these tasks seemed to lie in the lack of logical connections in the words presented, making it necessary to establish new associations.

Although there is substantial evidence that learning ability tends to decrease with advancing years, the results of research on retention have been more equivocal (Wimer, 1960; Wimer & Wigdor, 1958). Not only have studies on relearning and recall yielded conflicting results but investigators have been criticized for not controlling for original learning, an omission which made it difficult to determine whether older *S*s were poor learners, had poorer memories, or both.

One explanation offered to account for the relatively less adequate performance of the aged is a decline in motivation to learn. Donahue (1956) observed that attitudinal and emotional factors which appear in the later years of life have a deleterious effect on learning. Welford (1951) found that older persons have shown increasing insecurity under conditions of threat. Welford noted that *S*s over 30 were reluctant to submit to psychological testing and that even if they could be persuaded to participate in a learning task, they frequently revealed anxiety concerning the quality of their performance. Donahue (1956) has suggested that the manner in which a task is presented to an individual may be an important factor in eliciting good or poor performance. Wimer (1960) proposed that inasmuch as previous research has concentrated largely on the effect of variation of the material to be learned, future investigators should include variation in the kinds of instructions given the learner.

Although no experimental work has been done on changes in performance as a function of motivational instructions in aged persons, considerable interest has

been directed toward the combined effects of experimentally induced verbal stress and measured anxiety in college students (Sarason, 1956; Sarason, 1957; Sarason, 1958; Sarason & Palola, 1960). Performance levels have been observed to change if the individual was told that a test was a measure of intelligence or that his score would be compared with that of others.

The purpose of the present study was to determine whether age-related differences in paired-associate learning and retention could be modified by the experimental manipulation of two motivational variables. It was hypothesized that (*a*) the uncommon paired-associate list, representing the more novel and difficult material, is acquired and relearned in a greater number of trials by elderly persons under challenging instructions and (*b*) the difference in performance between young and old on the identical task would be smaller under supportive than under challenging or neutral experimental treatments.

METHOD

Subjects

One hundred twenty noninstitutionalized American-born white males, half of whom were between the ages of 18 and 26 and the remainder within the 65 to 75 year age range were used for the study. Individuals with severe auditory or visual limitations, diabetes, a history of alcoholism, cerebral vascular accidents, or central nervous system insult were excluded from the sample. The majority of *S*s were drawn from social and recreational clubs throughout greater New York. The elderly *S*s were retired individuals who were active members of day centers. The majority of young *S*s were employed in either part- or full-time jobs.

All *S*s were tested on verbal ability as determined by the Gallup Thorndike Vocabulary Test (GT Test) Form 1 and 2 (Thorndike & Gallup, 1944). Only those *S*s who achieved a mean raw score of correct responses between 7.5 and 14.5, representing the 25th to 75th percentile of the Gallup Thorndike Voting Sample of 1944, were included in the study. Socioeconomic status was evaluated on the basis of Reiss' (1961) socioeconomic index.

Learning Materials

The learning tasks were two lists of paired-associate nouns of 10 pairs each. The stimulus words, selected from the Kent-Rosanoff 100 Word List (Russell & Jenkins, 1954), were identical for both lists. The response items differed in level of difficulty (Ross, 1967). One list contained pairs which, according to a pilot study conducted by the experimenter (*E*) were found to be relatively easy. The pairs were selected on the basis of an associative strength between 11% to 55% as determined by Tresselt's[1] norms. These norms were based on the responses and the frequency of responses given to the stimulus words by groups comparable to those used in the present study. The associative strengths of the words chosen were equivalent for both age groups.

The response words representing the more difficult words were selected on the basis of 0% to .8% associative strength. For example, a response term of a paired associate of zero associative strength never appeared as a response to a stimulus word on Tresselt's norms, while an associative strength of .7% or .8% represented only one response to a stimulus word given by her younger and older *S*s, respectively. No words were included in the two lists which were beyond the eighth grade difficulty level on the Thorndike-Lorge (1951) lists. The easy paired associates have been designated "common" and the more difficult paired associates, "uncommon," throughout the experiment. The stimulus and response terms were printed in black letters ½-inch high × ⅛-inch wide on 3 × 5 inch unlined white index cards. On one side of the card appeared a stimulus word, and on the reverse side the response word. An electric timer, placed out of *S*'s view, was set to flash on and off for 1 second at 5 second intervals.

Procedure

Three to 5 weeks before *S*s participated in the experiment, each *S* was administered the GT Vocabulary Test. No *S* was informed of the purpose of the test but was merely told that the psychologist was interested in finding out which words were considered by most people to be easy and which were considered hard.

Those *S*s who had fulfilled the vocabulary criterion were sent a letter asking them to participate in a research project to be held at the day or recreation centers. The 120 volunteers were assigned randomly to one of six experimental groups as follows: Twenty of the old and 20 of the young *S*s comprising Groups 1 and 4 were given neutral instructions; 20 of the old and 20 of the young *S*s comprising Groups 2 and 5 were given supportive instructions; 20 of the old and 20 of the young comprising Groups 3 and 6 were given challenging instructions.

Each *S* was tested individually by *E*. After *S* was seated, he was given either the neutral, supportive, or challenging instructions as follows:

Neutral Instructions

I am going to show to you and read to you a list of words, two at a time. When I finish read-

[1] Margaret Tresselt, personal communication, October 2, 1963.

ing the two words that go together, I am going to say one word of each pair and ask you to tell me the word that went with it. For example, if the words are EAST-*west*, GOLD-*silver*, then when I say the word EAST, I would expect you to say (pause) *west*. And when I say the word GOLD, you would, of course, answer (pause) *silver*. Do you understand?

These instructions were repeated for the groups receiving the supportive and neutral instructions after the reference to Columbia University.

Supportive Instructions

I need your help with a research project that I am doing for Columbia University.... I am interested in finding out something about the characteristics of words. Your performance is *not* my main concern. My purpose in asking you to do this task is just to find out which words go together more easily and which do not.

Challenging Instructions

I am doing a research project for Columbia University.... The ability to learn this material is a good test of your intelligence, not of what you know but of how well you can learn *new* things. It's to your advantage, then, to do your best to show how capable you are, how bright you are in relation to people of your own age. Listen carefully and do your best, for your score will be compared with those of other subjects.

Acquisition of Paired Associates

The paired-associate lists were presented to *Ss* by the anticipation method. The order of presentation of the lists was alternated so that half of the *Ss* received the common pairs first, and the balance of the *Ss*, the uncommon pairs first. The *E* held a card in front of the seated *S* and called out the stimulus word. After a 5-second interval, the card was turned to expose the response term for 5 seconds. When the entire list had been presented in this manner, another 5-second interval was allowed while the cards were shuffled. On the next trial the stimulus word was again shown to *S* while *E* called the stimulus card as it appeared, but on this, and subsequent trials, *S* was required to give the response term. If the correct association was supplied, *E* repeated the word, reversed the card and proceeded to the next pair after a 5-second interval. If *S* made an error or failed to give any response, *E* supplied the correct association. The *Ss* who did not reply within the allotted time were given an additional 5 seconds during which time the cards were reshuffled to randomize the order of presentation of the word pairs. This procedure was repeated until a criterion of two correct recitations of the list was met. The relatively slow rate of presentation was selected to allow sufficient time for the elderly *Ss* to make a response. Persons who failed to learn the list within 30 trials were excluded from the study.

During the ½ hour interval between the acquisition and relearning of the paired-associate list, *S* and *E* were engaged in working on simple jigsaw puzzles.

Relearning of Paired Associates

The *E* introduced the relearning part of the experiment as follows: "I am going to read to you and show to you the *same* words that you had before. As you recall, I shall give you a word and you will give me the word that goes with it." The same instructions as given before the learning trials were repeated except that the reference to Columbia University was omitted. The procedure continued as before. Upon termination of the task, *S* was dismissed and told to return in ½ hour to continue the experiment.

The *Ss* spent the interval before the learning and relearning of the second list pursuing their usual activities in the day center. At the stipulated time *Ss* were tested for their learning of the second list. The *E* began the session by saying, "Here are some more words that I am going to show to you and read to you...." The exact procedure as outlined for the presentation of the first list was followed for the acquisition and relearning of the second list. An anxiety self-rating scale, devised by *E*, was administered at the end of the experimental session to enable *Ss* to rate themselves on a 5-point continuum according to the degree of tension and anxiety experienced during testing. Upon completion of the scale, an autobiographical questionnaire was given to *Ss*. Several weeks after all *Ss* were tested, five psychologists evaluated 12 recorded experimental sessions of the elderly *Ss*, equally divided among the three treatments.

RESULTS

Before analyzing the results, the distribution of relevant control variables was examined. Separate 2×3 analyses of variance for age and type of instruction were computed for socioeconomic status scores, GT vocabulary scores, anxiety self-rating scores, and number of years of schooling. The F ratios were not significant for socioeconomic status, GT vocabulary, and anxiety self-rating. However, the F ratio for education was significant ($F = 54.10$, $p < .001$). There were no significant interaction effects for any of the variables.

Additional analyses were made to determine the extent to which differences in education might be related to learning trials. When education was correlated with acquisition scores for old and young separately, the correlation was not significant

TABLE 1

MEANS, STANDARD DEVIATIONS, AND DIFFERENCE SCORES OF NUMBER OF TRIALS FOR LEARNING COMMON AND UNCOMMON PAIRED-ASSOCIATES TO CRITERION BY YOUNG AND OLD GROUPS UNDER THREE EXPERIMENTAL TREATMENTS

Experimental Treatment	Common					Uncommon				
	Young		Old		Difference	Young		Old		Difference
	$\bar{X}$	*SD*	$\bar{X}$	*SD*		$\bar{X}$	*SD*	$\bar{X}$	*SD*	
Neutral	2.20	.40	3.30	1.14	1.10	5.80	2.44	16.40	5.32	10.60
Supportive	2.35	.57	2.95	1.00	.65	5.20	2.02	13.20	4.99	8.00
Challenging	2.40	.26	3.35	1.10	.95	5.50	2.06	22.00	6.93	16.50

Note.—N = 20 in each subgroup.

for either group (.04 and .20, respectively). Thus, within each age group, education bears no relationship to efficacy of performance. However, a correlation of −.44 between education and learning trials for young and old *S*s combined was significant. Age also correlated significantly with acquisition trials ($r = .75$). A partial correlation between age and learning trials, eliminating the effects of education, was run and the results showed that age and performance were correlated .68. This high correlation probably owes much of its magnitude to the wide age spread between the young and old groups. When the influence of age was partialed out, education correlated only .03 with acquisition trials, indicating that age was the factor most responsible for the differential performance between the groups.

Under the challenging treatment older *S*s required approximately one-third more trials to fulfill the learning criterion for the uncommon words than under either of the other conditions (Table 1). A simple analysis of variance showed that for old *S*s the difference between the experimental groups was highly significant ($F = 11.17$, $df = 2/57$, $p < .001$). In addition to the significant F ratio, t ratios were calculated to determine more specifically the locus of differences in the experimental groups. A comparison of the mean scores of the older *S*s between the neutral versus challenging and the supportive versus challenging instructions yielded significant t values of 2.8, $p < .01$ and 4.6, $p < .001$, respectively. The performance of older people appeared to be at its best in a supportive situation, but at its worst in a challenging one.

The older *S*s not only required more trials to reach the learning criterion but displayed greater variability in their performance than did the younger *S*s. Difference scores were obtained by subtracting the mean number of trials required by the younger *S*s from the mean number of trials required by the older *S*s to reach the learning criterion. Table 2 summarizes the t ratios of the differences between the difference scores of all *S*s under the three experimental conditions. The difference between young and old in the number of trials needed to reach the criterion under the supportive instructions was significantly smaller than under the challenging treatment, $t = 4.17$, $p < .001$. However, when the difference between the performance of both groups under the challenging and neutral conditions was analyzed, the

TABLE 2

DIFFERENCE SCORES BETWEEN YOUNG AND OLD SUBJECTS UNDER NEUTRAL, SUPPORTIVE, AND CHALLENGING TREATMENTS FOR ACQUISITION OF UNCOMMON PAIRED-ASSOCIATES

Treatment	M_{diff}	*SD*	*diff*	*t*
Neutral	10.6	5.1		
			2.6	1.62
Supportive	8.0	5.1		
			8.5	4.47*
Challenging	16.5	7.9		

Note.—N = 20 in each subgroup.
* $p < .001$.

value of t did not approach significance at the 5% level.

The elderly *Ss* took approximately twice as many trials to relearn the uncommon list as the younger *Ss* and also had larger standard deviations over all treatments. However, the differences among experimental treatments were not significant ($F = 1.36$). There was no significant effect for treatment on the relearning phase of the experiment, nor was there an interaction between age and treatment. Unlike their effect upon original learning, challenging and supportive instructions did not appear either to depress or to facilitate performance on the more difficult tasks.

When the difference scores between the two age groups for relearning the uncommon pairs under the three experimental treatments were compared, there was no significant differential effect as a function of instructions.

Although there was a significant difference between young and old in the mean number of trials needed to learn common paired associates ($F = 15.64$, $p < .001$; 4.97, $p < .05$; 10.28, $p < .01$) under neutral, supportive, and challenging instructions, respectively, there was no differential effect as a function of treatment. Despite the fact that the elderly group took longer than the young to master the easier task, the performance of older *Ss* was far more adequate on the common than on the uncommon list. There was no significant age-related difference for the common pairs on the relearning part of the experiment.

DISCUSSION

Elderly *Ss* showed a decrement in performance in acquiring the uncommon pairs under challenging instructions on the more difficult task. The increasing insecurity and susceptibility to stress of aging individuals become particularly evident when they are placed in an evaluative situation and told that their performance will be compared with that of others. Research in anxiety and verbal learning indicates that the efficiency of *Ss* who score high on an anxiety scale seems to be more disrupted by high motivation or personal threat situations than the performance of individuals with lower scores (Sarason, 1960). Sarason (1958) noted that anxiety appears more effective in depressing performance when the material to be learned is difficult or complex. He also observed that under neutral or reassurance conditions differences in performance between *Ss* differing in test anxiety have been negligible or reversed, a finding which is consistent with the result in the present study that elderly *Ss* did best under supportive and worst under the challenging instructions.

Although the older group required significantly more trials than the young to relearn the more difficult material, they did not appear to be differentially affected by the instructions. Obviously the challenging treatment did not have the disruptive effect upon performance on this phase of the experiment that it did when the material was being acquired. It is possible that when the material was presented for the second time it may no longer have seemed novel nor as difficult as it did initially. Another explanation may be that with increased familiarization the older *Ss* learned to reduce anxiety through successful performance of the task in the earlier phase of learning.

While the elderly *Ss'* performance on the relearning of the uncommon list was poorer than that of the younger *Ss* under all treatments, proportionate to their acquisition, their rate of relearning was better than that of the young. This seeming superiority of the older group may be attributed to the availability to them of more ceiling for improvement.

A qualitative assessment of *Ss'* behavior during testing made by *E*, in addition to the Anxiety Self-Rating scale, showed that a larger proportion of the older *Ss* under the challenging than under the neutral and supportive conditions made comments as to their awareness of increasing memory loss and "stupidity." Six of the older *Ss* under the challenging treatment as compared with three under the neutral,

and none under the supportive instructions failed to respond correctly to any items on the first learning trial. Although it cannot be presumed that anxiety per se was responsible for the discrepancy in age groups, it does appear that the older person reacts with nonadjustive behavior in an evaluative situation where his performance is to be compared with that of others.

The findings firmly support empirical evidence as to age-associated decrements in learning. However, motivational variables appeared to be most effective in the initial stage of learning. If performance can be modified in the aged, the implications for future work with older persons are manifold. Just as new teaching methods have facilitated learning in children, so new approaches to training and rehabilitation may result in narrowing the performance gap between the generations.

REFERENCES

DONAHUE, W. Learning, motivation, education of the aging. In J. E. Anderson (Ed.), *Psychological aspects of aging*. Washington, D. C.: American Psychological Association, 1956.

GILBERT, J. G. Memory loss in senescence. *Journal of Abnormal and Social Psychology,* 1941, **36,** 73–86.

KORCHIN, S. J., & BASOWITZ, H. Age differences in verbal learning. *Journal of Abnormal and Social Psychology,* 1957, **54,** 64–69.

REISS, A. J., JR. *Occupations and social status*. New York: Free Press of Glencoe, 1961.

ROSS, E. Effects of challenging and supportive instructions on verbal learning in older persons. Unpublished doctoral dissertation, Columbia University, 1967.

RUSSELL, W. A., & JENKINS, J. J. *The complete Minnesota norms for responses to 100 words from the Kent-Rosanoff word associates list.* Technical Report No. 11, 1954, Contract No. 8, ONR-66216, Office of Naval Research.

SARASON, I. G. Effect of anxiety, motivational instructions, and failure on serial learning. *Journal of Experimental Psychology,* 1956, **51,** 253–260.

SARASON, I. G. Effect of anxiety and two kinds of motivating instructions on verbal learning. *Journal of Abnormal and Social Psychology,* 1957, **54,** 166–171.

SARASON, I. G. Effects on verbal learning of anxiety, reassurance, and meaningfulness of material. *Journal of Experimental Psychology,* 1958, **56,** 472–477.

SARASON, I. G., & PALOLA, E. G. The relationship of test and general anxiety, difficulty of task, and experimental instructions to performance. *Journal of Experimental Psychology,* 1960, **59,** 185–191.

THORNDIKE, E. L., & LORGE, I. *The teachers' word book of 30,000 words.* New York: Columbia University, 1951.

THORNDIKE, R. L., & GALLUP, G. H. Verbal intelligence of the American adult. *Journal of Genetic Psychology,* 1944, **30,** 75–85.

WELFORD, A. T. *Skill and age*. London: Oxford Univer. Press, 1951.

WIMER, R. E. A supplementary report on age differences in retention over a twenty-four hour period. *Journal of Gerontology,* 1960, **15,** 417–418.

WIMER, R. E., & WIGDOR, B. T. Age differences in retention of learning. *Journal of Gerontology,* 1958, **13,** 291–295.

(Received July 18, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 4, 267–274

PERSONALITY AND MOTIVATIONAL FACTORS IN RESPONSES TO AN ENVIRONMENTAL DESCRIPTION SCALE

EDMOND MARKS
Georgia Institute of Technology

The effects of 2 classes of variables—(*a*) selected personality and motivational factors, and (*b*) certain task parameters in the form of item content and item ambiguity—on judgments by college age students of their college environment were investigated. Contrary to the assumptions implicit in most "perceptual" environmental assessment scales, perceptions and thus descriptions of the college environment were not independent of the properties of S—e.g., personality and motivational characteristics—or properties of the items. Furthermore, it was found that these 2 classes interacted to yield differential effects on environmental judgments. The results were discussed in terms of the role which properties of S play in the handling of ambiguous or inadequate informational input to maintain a stable and expected cognitive frame.

In recent years a considerable portion of educational research has been concerned with the study of environmental or situational determinants of observed educational behaviors. While most of this research has been limited to the college environment, the rationale and procedures employed with the college setting are readily generalized to other environments and other behaviors. Numerous college environmental assessment techniques have been developed which purport to measure in some sense the dominant characteristics of the college environment. One popular class of these techniques concerns student perceptions and cognitions of the college environment, perhaps the most notable being Stern's (1958) College Characteristics Index (CCI) and Pace's (1963) College and University Environment Scales (CUES). In using scales like these the typical procedure is to average responses from a set of respondents who are considered homogeneous with respect to some characteristic of interest, and report these mean values as describing or profiling the particular environment studied.

If a given environment is treated as a constant set of stimuli at any given point in time, then variability in response to this set of stimuli should be attributed to random error or to selected characteristics of the respondent or items. Foremost among such characteristics of the respondent one might specify stylistic variance—such as responding in a socially desirable or acquiescent manner—reliable personality differences, or differences in the perception of and the meaning attributed to the stimuli. Alternatively, it may be that with a complex environment the respondents are not all attending to the same elements within the set. Inequality of subsets of cues, even if they overlap, could lead to different responses to the same item. Whichever the case, for purposes of profiling an environment the variability about the expected value must be treated as error variance which is to be minimized for effective assessment (Torgerson, 1962).

To indicate the extent of this variability, in one sample of 611 fall term 1965 Georgia Institute of Technology freshmen, the range of item variances for the 300 items of the CCI was .009–.250, with the maximum possible being .250. The median item variance was .167. Since the variance of any scale is a function of the variances and intercorrelations of the items comprising that scale, this rather large median item variance suggests considerable variability in scale scores, and thus considerable lack of uniformity in assessing characteristics of the environment. Pace (1963) aptly described this case when he said "... what is really characteristic of the school is that the students disagree about its characteristics! [p. 37]"

Where substantial item variances obtain on a college environment scale, a major problem is encountered in deciding what is the most appropriate method of scoring the scale. A more important and intriguing question, one affecting both the substantive development and construction of such scales, concerns whether this variance can be accounted for by selected characteristics of the respondents and the items themselves, rather than the environmental aspect sampled.

Two item characteristics of immediate appeal in this respect are item content and item ambiguity. Items comprising most "perceptual" environmental description scales vary widely along a continuum from high to low cue determinancy. Some items appear unambiguous in that they are easily verified by scanning the environment, for example, "there are no fraternities or sororities." Responses to items describing an ambiguous aspect of the environment—for example, "personality, pull, or bluff get students through many courses"—seem particularly susceptible to the effects of certain perceptual or personality processes. For example, a student who is abasing, dependent, and fearful of his academic performance, regardless of his ability, may be quite defensive in response to items like the one cited relating to getting through courses.

Another relevant aspect of this problem is the familiarity or sampling dimension. Extrapolating from perceptual learning studies (Wohlwill, 1966) one might expect a change in college student perceptions toward increased veridicality as a result of increased sampling or familiarity with the environment. Again, however, certain properties of the individual might be tied to parameters of this function as when some students are more resistant to perceptual change under conditions of increasing input than are others. Associated with environmental unfamiliarity, as would probably be most true of the entering college freshman, one might expect greater response uncertainty and the more pronounced influence of personality, motivational, and attitudinal factors upon item response (Cronbach, 1950; Gage, Leavitt, & Stone, 1957).

These comments form the background for the hypothesis being examined in the present study. Simply stated it is that a significant portion of what is presently assumed to be random error variance in scores on a selected college environment scale like the CUES can be attributed to the nonrandom effects of personality and sampling processes as they are elicited by selected item characteristics. In particular, it is postulated that item ambiguity and item content are reliably related to item variance, and that under certain conditions, for example, high item ambiguity, these item characteristics lead to the increased effects of selected personality and sampling variables upon item response.

An important question encountered in testing this hypothesis is "... on what basis should item ambiguity be defined?" One possibility is to define it in terms of expert judgment of the visibility of the environmental characteristic represented by the item. Another possibility is to define item ambiguity as the uncertainty experienced by *S* in making an item response. The first definition clearly reflects the inherent or "ideal" ambiguity of an environmental cue, that is, that ambiguity of a cue that remains even after the respondent has had some experience with the environment. The latter definition is more sensitive to the perceptual and judgmental processes peculiar to the individual respondent, and indicates, to a greater extent, his unique sampling tendencies. These two approaches to defining ambiguity differ also in an important theoretical respect (Spence, 1944). Ambiguity defined as response uncertainty is primarily a response-inferred construct being tied almost entirely to *S*'s responses, whereas expert judgment is more nearly a stimulus characteristic. In this study, a measure of each approach was included; the measure based on expert judgments being called "judged item ambiguity," while the measure based on student certitude responses was called "item-response certitude."

The evidence relating to *S* correlates of responses to perceptual environmental scales

is somewhat sparse (Herr, 1965; McFee, 1961; Saunders, 1962). After an extensive factor analysis of the CCI and its companion personality scales, the Activities Index (AI), Saunders concluded that the scale scores of the environmental measure were independent of the personality of the respondent. This conclusion was based upon the finding that, in general, the vectors defining each index spanned a unique subspace of the total factor space. That is, the total number of interpretable factors obtained by factoring the CCI and AI together could be broken down into two sets of factors, one group "that are loaded mostly by CCI variables and a second group ... loaded primarily by AI variables" (Saunders, 1962, p. 8). These data were not, however, completely "clean," there being some confounding of factor structures. Using the same scales, McFee arrived at the same conclusion. In contrast, Herr in studying the High School Characteristics Index obtained significant relationships between scores on this measure and certain ability and biographical variables. Also of interest for the present study was Saunders' and Herr's finding of a rather substantial error variance for the environmental measures employed.

METHOD

Measures Employed

Environmental measure. The college environmental scales selected for study were the CUES (Pace, 1963). This inventory consists of 150 items which are broken down into five nonoverlapping scales of 30 items each. The five scales, labeled practicality, awareness, community, propriety, and scholarship were defined on the basis of a factor analysis of the intercorrelations among the means of the 30 CCI scales for a sample of 50 colleges and universities. Items for the CUES were then selected from the 300 CCI items in terms of how well a given item defined one of the five scales.

Subject variables. The personality and motivational variables used in this study were selected in terms of their hypothesized relationships with the content of the CUES items as defined by Pace (1963). Nine of the personality scales were drawn from the 22 scales of Jackson's (1965) Personality Research Form, Form A, each of which contains 20 items. The scales selected were achievement, affiliation, autonomy, cognitive structure, dominance, order, social recognition, succorance, and understanding. The test-retest reliabilities reported by Jackson (1965) for these nine scales ranged from .73 for cognitive structure to .88 for dominance. Two 10-item scales developed by Marks and Messersmith (Marks, 1967) relating to motivational aspects of educational behavior were also included. The two scales were level of educational and career aspiration and fear of failure, whose split-half reliabilities were .89 and .80, respectively.

To evaluate differences due to cue sampling, the student was asked to indicate on a 5-point scale, the amount of information he had about, or how familiar he felt he was with, the college environment. Finally, each student was asked to indicate again on a 5-point scale, how certain he was when responding to a given CUES item, of the accuracy of that response. This variable was referred to as "item-response certitude" in the analysis.

Procedure

Item ambiguity and content definitions. The 150 CUES items were independently rated for ambiguity by five trained psychologists. The raters were asked to place the items into five ordered categories on the basis of the "extent to which the item reflected a characteristic of the environment which was difficult to verify perceptually or for which the stimulus cues would tend to be vague, subtle, or conflicting." Each rater was asked to read the entire list of items once before rereading them for categorizing purposes. Items for which there was less than a 4 to 1 agreement were deleted from that part of the analysis relating to ambiguity. This variable was denoted "judged item ambiguity."

The 150 CUES items were grouped according to content in terms of the five first-order factors reported by Pace (1963). The content of an item was defined simply by Pace's description of the scale to which the item belonged.

Several related analyses were conducted on the 15 S and item variables, and the dependent variable—the proportion endorsing each item. Before carrying these out, however, the dependent variable was transformed in order to more clearly reflect the parameter of interest, that is, item variance. Since the major concern in this study was with item variances and their correlates, the proportion endorsing each item was transformed so that all p values fell within the range $.50 \leq p \leq 1$. This was accomplished by setting $p \geq .50$ equal to p, and $p < .50$ equal to $1 - p$. Since the mean and variance of the binomially distributed CUES items are inversely related under this transformation, p values tending towards .50 indicate increased item variance.

The item parameters—ambiguity, content, response certitude, and proportion responding—were intercorrelated and their means and standard deviations computed. Those correlation estimates involving item content represent contingency coefficients, while all others are product-moment correlations. In addition, the means, standard deviations, and intercorrelations among the 12 S variables were computed.

Examination of the joint distributions of a given CUES item and the *S* variables indicated that, in many cases, the variables did not form a bivariate normal density. Because of this condition it was decided to examine the relationship of CUES item response to the selected *S* variables by testing for differences between the cumulative distributions of a single *S* variable for the two item-response categories—true or false. Should a relationship exist, scores for one of the categories would be expected to shift toward higher values. For this purpose a test due to Kolmogorov and Smirnov was used (Siegel, 1956).

As previously indicated, only those personality and motivational variables which were suspected on the basis of the congruence of variable and item contents of being sensitive to the hypothesis of a significant item-variable correlation were selected for study of their relationship to a given CUES item. Scholarship items were related to achievement, level of aspiration, and fear of failure; propriety items to cognitive structure, dominance, and order; community items to affiliation, autonomy, and succorance; practicality items to affiliation, order, and social recognition; and finally, awareness items to the single variable of understanding. The items were also related to reported familiarity with the environment.

To facilitate the analysis, the number of CUES items examined was reduced by systematically selecting a smaller number of items which would permit evaluation of the hypothesis of an interaction of item-*S* parameters. The 150 CUES items were cross-classified in a 6 × 5 table defined by item content and mean item-response certitude, and a total of 25 items selected by randomly choosing one item from each cell. Since the two item parameters were correlated in the sample, sampling of the items was not uniform over the 30 cells of this table.

Subjects

The *S*s were 570 male freshmen entering Georgia Institute of Technology in the fall term of 1966. The tests were administered during the week prior to registration.

RESULTS

The means, standard deviations, and intercorrelations of the four item variables are presented in Table 1. The values involving item ambiguity are based on only 139 cases. Eleven items had to be deleted because they failed to satisfy the criterion of a 4 to 1 agreement among judges. Because of differences in the correlational methods employed, conclusions are best limited to statements concerning significance, not magnitude.

TABLE 1

MEANS, STANDARD DEVIATIONS, AND INTERCORRELATIONS AMONG p, AMBIGUITY, CERTITUDE, AND CONTENT

Item	Ambiguity	Certitude	Content[a]	*N*	*M*	*SD*
p	−.05 (*ns*)	.58*	.42*	150	74.8	17.3
Ambiguity		−.08 (*ns*)	.28 (*ns*)	139	2.9	1.2
Certitude			.47*	150	3.5	.6

Note.—Since the Content classification had no ordinal properties, the mean and standard deviation are not reported.
[a] Contingency coefficients.
* $p < .05$.

In spite of these qualifications it is apparent that there is no association between judged ambiguity of an item and the three other item characteristics. The hypothesized correspondence between judged item ambiguity and the students' cognitive and response processes, that is, certitude judgments and the proportion selecting a given alternative, failed to emerge. Item content, on the other hand, was significantly correlated with both these cognitive and response processes. An inspection of the respective contingency tables indicated that these correlations were due primarily to two of the five content categories; scholarship and awareness. Scholarship items tended to have high item-response certitude means and high p values or low item variances, while the awareness items tended to have low mean response certitude values and p values which tended more toward .50—high item variances.

Mean item-response certitude, as an index of the indeterminancy the item possessed for the sample of students, aside from correlating significantly with item content, also correlated substantially with the proportion selecting a given alternative. Items which were described by the sample as eliciting uncertainty as to the accuracy of response, tended to have high item variances.

The means, standard deviations, and intercorrelations among the 12 personality, motivational, and familiarity variables are presented in Table 2.

Although not of direct interest in terms of the hypotheses being examined, comment should be made on some of the correlations in Table 2. Quite noticeable is the lack of correlation between the students' reported familiarity with the college environment and the other variables studied. At least for this set of variables, students'

TABLE 2

MEANS, STANDARD DEVIATIONS, AND INTERCORRELATIONS AMONG THE 12 SUBJECT VARIABLES

Variables	Achievement	Affiliation	Autonomy	Cognitive structure	Dominance	Fear of failure	Level of aspiration	Order	Social recognition	Succorance	Understanding	Familiarity
Achievement		−.05	.09*	.04	.13*	.11*	.36*	.12*	−.01	.16*	.27*	.02
Affiliation			−.31*	.00	.12*	.10*	.03	.09*	.29*	.32*	−.02	−.03
Autonomy				−.13*	.10*	.13*	.08	−.06	−.20*	−.42*	.12*	.06
Cognitive structure					−.09*	.22*	.06	.51*	.03	.10*	−.10*	.08
Dominance						−.02	−.05	.03	.26*	−.02	.14*	−.01
Fear of failure							.13*	.26*	.01	.06	−.11*	.06
Level of aspiration								.03	.18*	−.02	.22*	.03
Order									.10*	.02	.03	.01
Social recognition										.19*	−.13*	.08
Succorance											−.14*	−.08
Understanding												.02
M	13.6	14.1	8.3	11.2	9.7	6.2	6.7	10.9	11.5	8.3	12.0	3.8
SD	3.5	3.5	3.0	3.2	4.3	1.4	1.8	4.1	3.6	4.1	3.4	1.2

Note.—$N = 570$, $df = 500$.
* $p < .05$.

judgments of their familiarity with the institution studied, were independent of the personality characteristics of the respondent.

The intercorrelations among achievement, level of aspiration, and understanding were suggestive of a form of investment in intellectual activity which has both motivational and cognitive components. This pattern is consistent with Murray's (1938) treatment of Need for Understanding, and perhaps, Tolman's (1951) "placing need." On the other hand, fear of failure, order, and cognitive structure were reliably correlated suggesting that students who are fearful of their performance tend to approach their personal and situational involvements in a cautious and orderly way, thus apparently reducing the perceived possibility of substandard performance. Students higher in these traits can be viewed as having difficulty in handling environmental situations which depart from the expected.

The tests of association between the selected CUES items—cross-classified on item content and mean item-response certitude—and the personality, motivational,

TABLE 3

SUMMARY OF THE TESTS OF ASSOCIATION BETWEEN THE SELECTED COLLEGE AND UNIVERSITY ENVIRONMENT SCALES ITEMS AND SUBJECT VARIABLES

Mean item-response certitude	Item content: Practicality	Awareness	Community	Scholarship	Propriety	% of tests significant
2–2.5		Understanding*				100
2.51–3	Affiliation Order Social recognition*	Undersatnding	Affiliation* Autonomy Succorance*		Cognitive structure Dominance* Order	40
3.01–3.5	Affiliation Order Social recognition	Understanding*	Affiliation* Autonomy Succorrance*	Achievement* Level of aspiration* Fear of failure*	Cognitive structure Dominance* Order	54
3.51–4	Affiliation Order Social recognition	Understanding*	Affiliation Autonomy Succorance*	Achievement* Level of aspiration* Fear of failure	Cognitive structure Dominance Order*	38
4.01–4.5	Affiliation Order Social recognition	Understanding	Affiliation* Autonomy Succorance*	Achievement* Level of aspiration Fear of failure*	Cognitive structure Dominance Order	31
4.51–5	Affiliation Order Social recognition	Understanding	Affiliation* Autonomy Succorance	Achievement Level of aspiration Fear of failure	Cognitive structure Dominance Order	8
% of tests significant	7	50	53	58	20	

Note.—$N = 570$.
$p < .05$.

and environmental familiarity variables are summarized in Table 3. Within each cell—corresponding to a single CUES item for a given level of mean item certitude and content class—the selected *S* variables are listed and the significance of the item-variable association noted.

The notion that personality and motivational variables are related to item response on an environmental assessment scale appears supported by the data; over 30% of the relationships tested were significant at the .05 level. This conclusion is offered cautiously, since the test criteria are probably not independent. In addition, the association between item response and the respective *S* variables appears to be moderated by the two item parameters studied. Scholarship items were, in most instances, significantly related to all three of *S* variables hypothesized to be relevant to this content class. Similarly, three of the six awareness items were significantly related to understanding, while community items appeared related to both affiliation and succorance. The results for the practicality and propriety classifications were, however, much less indicative of a reliable effect of personality and motivation upon responses to the CUES items.

Once again, familiarity failed to emerge as a correlate of response variability. Students' reports of their familiarity with the Georgia Institute of Technology environment bore little relation to the variability of their judgments of its characteristics.

Despite the caution noted concerning overall tests of significance, the results of this part of the analysis provide rather good evidence that responses to some items of the CUES are dependent upon certain characteristics of *S*s and the items.

DISCUSSION

The results of the present study, particularly those relating to *S* correlates of the CUES item variance, are perhaps best treated as providing reliable but limited evidence for the presence of nonenvironmental factors in the response to the items of a selected environmental assessment instrument. They are neither exhaustive of the possible relationships that might exist between these two domains, nor do they indicate the magnitude of the effects of such nonenvironmental variables upon item response. What these results do indicate is that for some of the selected *S* and item characteristics studied, a reliable portion of the response to a given environmental characteristic can be attributed to certain properties of *S*. Since it is rarely the intent of the constructor of environment scales to provide for an *S* component of the item variance, this reliable component must be incorporated in the error variance. For some CUES items what is really being characterized to a great extent is the sample of students—not the environment.

Of particular interest is the lack of association between the reliable judgments of the CUES item ambiguity and the students' reported item-response certitude. Students apparently develop, through some undefined mechanism, a set of stable perceptions and cognitions about the environment to which they are responding which is independent of the number and clarity of the environmental cues available. Given an item like, "There is a lot of apple-polishing around here," where one might suspect the environmental cues to be vague and poorly defined, one nonetheless, finds a very low endorsement value—$p = .04$. This raises the important questions of how are stimulus cues utilized by the student in making an environmental judgment, and second, how are these environmental perceptions and cognitions formed. Furthermore, although these perceptions and cognitions are consistent in that they are shared by the sample as a whole, there is the question of the veridicality of such judgments. It is doubtful whether items tapping an environmental aspect of high cue indeterminancy can reflect a uniform property of the environment, or lead to high consistency of response. As suggested later, a part of this response consistency, where environmental cues are vague or conflicting, might be attributable to selected personality and need structures of the student. A student who perceives his college academic environment as highly rigorous and demanding is unlikely to

engage in the dissonant response of endorsing the "apple-polishing" item, regardless of the nature and number of cues available on this environmental characteristic.

As hypothesized, certain item parameters were related to the variances of the CUES items, and equally important, they interacted with certain of *S* variables in determining response variability. In particular, certain of the factor analytically defined content classes were related to *S* uncertainty and item variability. This, in itself, may be a function of the institution or environment being studied. At Georgia Institute of Technology, for example, the emphasis upon academic achievement and competition, and the rigorous pursuit of the acquisition of knowledge is quite noticeable; this set of cognitions being shared by the students, faculty, and administration. This particular perceptual system—best described by Pace (1963) in terms of scholarship—provides considerable uniformity of response, high response certitude, high *p* values, and low item variances. On the other hand, the content area labeled awareness by Pace (1963), and described by him in terms of reflectiveness, self-understanding, interest in human welfare, and in general, a concern for "personal, poetic, and political" meaning, is much less clearly articulated at Georgia Institute of Technology. In this area the Georgia Tech student apparently has fewer and more poorly defined cognitions upon which to base his responses. In addition, the elements of this content class probably have less subjective utility for the Georgia Tech student during this interval of his life. The low response certitude mean values and response proportions tending towards .50 reflect this lack of a perceptual and cognitive frame with respect to this dimension.

As intimated earlier, responses to some items of the CUES are reliably related to selected personality and motivational variables, with these relationships being moderated by item content and the mean response certitude associated with the item. The environmental area where these effects were most pronounced was scholarship. Students high in achievement, level of aspiration, or fear of failure were considerably less variable in their response to items drawn from this content class than were students who obtained lower scores. Apparently, students who score high on these traits have a greater need to perceive their environment in a particular way and are more greatly affected by an environmental cue—in this case an item describing the environment—which is discrepant with their environmental expectations. This would be particularly true for items describing the environment which elicit considerable uncertainty among the respondents. In this case, the student—by not having a well articulated perceptual and cognitive frame for evaluating the given environmental cue—tends to rely more heavily on his personality or motivational domains as determinants of his CUES item response. It seems reasonable that under these conditions—that is, high environmental uncertainty and high need—the student will emit a CUES response that conforms to or is congruent with his particular need structure. As such, greater uniformity of response—that is, low item variances—may be viewed as reflecting an attempt by these students to maintain a congruence between the kinds of environmental supports they seek, and their perceptions of certain environmental inputs as implied in the items. Under this interpretation, for example, a student high in achievement who perceives himself as a hard worker who earns everything he gets, is less likely to engage in a dissonance-producing response of endorsing an "apple-polishing" or "personality, pull, and bluff" type item, or any other item involving achievement by means of duplicity.

A similar interpretation, utilizing congruence between personality or need states and environmental input as the mechanism underlying item response, can be offered for the other content classes where significant associations were obtained. Here again the perceptual needs of the students as defined by their scores on the personality and motivational variables are left intact by the highly selective response to the environmental items. This

interpretation is obviously consistent with and draws heavily from the dissonance and imbalance theories of Festinger (1957), Heider (1958), and others.

To this point, the authors have been stressing the role of item content as moderating the relationship of the selected personality and motivational variables to item response. Mean item-response certitude also tends to serve a similar function, although its effects are less pronounced than item content. A part of this attenuation may be tied to the increasingly restricted range of CUES item responses as mean item-response certitude increases. Nonetheless, both the certitude an S ascribes to the accuracy of his item response and item content are related to the nature and magnitude of the correlations between the personality and motivational variables, and item response.

A final statement should be made concerning the relationship between the magnitude of uncontrolled or error variance in item scores, and the method of factor analysis employed in constructing and interpreting the CUES scales. The irrelevant personality and motivational factors demonstrated in this study, serve to increase both item and scale variances. With scale scores defined on a fixed interval—in this case from 0 to 30—this increased scale variance has the effect of pulling the institution means closer together. Although factoring group means would appear to disregard S differences, focusing rather upon institutional differences, it is apparent that S differences reemerge by attenuating group mean covariances. Unreliability of the group scale means—as reflected in the scale variances and differentially contributed to by the personality and motivational factors—must be considered when using a procedure like that employed in constructing the CUES.

REFERENCES

CRONBACH, L. J. Further evidence on response sets and test design. *Educational and Psychological Measurement,* 1950, **10,** 3–31.

FESTINGER, L. *A theory of cognitive dissonance.* Stanford: Stanford University Press, 1957.

GAGE, N. L., LEAVITT, G. S., & STONE, G. C. The psychological meaning of acquiescence set for authoritarianism. *Journal of Abnormal and Social Psychology,* 1957, **55,** 98–103.

HEIDER, F. *The psychology of interpersonal relations.* New York: Wiley, 1958.

HERR, E. L. Differential perceptions of "environmental press" by high school students. *Personnel and Guidance Journal,* 1965, **43,** 678–686.

JACKSON, D. N. *The Development and evaluation of the Personality Research Form.* London, Canada: Author, 1965.

MARKS, E. Student perceptions of college persistence and their intellective, personality, and performance correlates. *Journal of Educational Psychology,* 1967, **58,** 210–221.

McFEE, A. The relation of students' needs to their perceptions of a college environment. *Journal of Educational Psychology,* 1961, **52,** 25–29.

MURRAY, H. A. *Explorations in personality.* New York: Oxford University Press, 1938.

PACE, C. R. *College and University Environment Scales: Preliminary manual.* Princeton: Educational Testing Service, 1963.

SAUNDERS, D. R. *A factor analytic study of the AI and CCI.* Princeton: Educational Testing Service, 1962.

SIEGEL, S. *Nonparametric statistics.* New York: McGraw-Hill, 1956.

SPENCE, K. W. The nature of theory construction in contemporary psychology. *Psychological Review,* 1944, **51,** 47–68.

STERN, G. G. *College Characteristics Index: Preliminary manual.* Syracuse: Author, 1958.

TOLMAN, E. C. A psychological model. In T. Parsons and E. A. Shils (Eds.) *Toward a general theory of action.* Cambridge, Mass.: Harvard University Press, 1951.

TORGERSON, W. S. *Theory and methods of scaling.* New York: Wiley, 1962.

WOHLWILL, J. F. Perceptual learning. In P. R. Farnsworth (Ed.) *Annual review of psychology.* Palo Alto: Annual Reviews, 1966.

(Received July 20, 1968)

Journal of Educational Psychology
1968, Vol. 59, No. 4, 275–282

Source and Direction of Causal Influence in Teacher-Pupil Relationships[1]

ALBERT H. YEE
University of Wisconsin, Madison

To study the source and direction of influence in teacher-pupil relationships, attitudes of teachers and their intermediate-grade pupils were measured early in September and several months later. Sample included 102 teachers with middle-class (MC) pupils and 110 teachers with lower-class (LC) pupils. The MTAI and a semantic differential measured teachers' attitudes; a 100-item About My Teacher inventory measured pupils' attitudes. Factor analyses provided homogeneous measures. Hypotheses dealing with teacher vs. pupil influence toward congruity and incongruity were tested, with *Ss* differentiated by teachers' experience and pupils' social class. Analyses of source and direction of attitude shift with the FCP technique showed: (*a*) influence of teachers' attitude predominates over LC pupils; (*b*) MC pupils and their teachers have more mutual influence relationships; (*c*) greater differences between levels of pupils' social class than between levels of teacher experience; and (*d*) serious need to improve teachers' relationship with LC pupils.

Teachers differ in their attitudes of warmth and permissiveness toward pupils. Classes differ in the favorability of pupils' perceptions toward their teachers. The two kinds of differences have consistently been found to correlate about .2 to .6 in the upper elementary grades (Getzels & Jackson, 1963, pp. 508–522); that is, warm teachers tend to be found in classes whose pupils like their teacher.

When a correlation occurs, the question of causality may be raised. In teacher-pupil attitude relationships, do the teacher's attitudes cause the class' favorability toward their teacher? Or does it work the other way around, so that friendly pupils make the teacher become warm and permissive? Theories and studies of social interaction portray the teacher-pupil relationship as complex and reciprocal (Bush, 1954; Della Piana & Gage, 1955; Flanders, 1965; Gage, Runkel, & Chatterjee, 1963; Ryans, 1960; Smith, 1960; White & Lippitt, 1960; and Withall & Lewis, 1963, pp. 708–710). Thus, the direction of influence in teacher-pupil relationships merits investigation.

Earlier studies (Heil & Washburne, 1962; Hoyt & Cook, 1960; Rabinowitz & Rosenbaum, 1960) have indicated that teachers' attitudes of warmth and permissiveness vary with years of teaching experience. Preservice and beginning teachers' scores on the Minnesota Teacher Attitude Inventory (MTAI) are considerably higher, on the average, than those of teachers with some experience (Beamer & Ledbetter, 1957). After 2 years of teaching, MTAI scores become stabilized at about the level found prior to teacher preparation. Such change in attitudes seems to result from interaction with pupils and not merely from the passage of time; thus, Day (1959) found that graduates who prepared for but did not enter teaching shifted less in attitudes than did those that entered teaching.

From this evidence concerning the reciprocity of teacher-pupil relationships, the decline of warmth and permissiveness during the first years of teaching, and the

[1] Appreciation is expressed to N. L. Gage who gave freely of his advice and encouragement throughout the project. The research reported herein was performed pursuant to contract OE-6-10-077 with the United States Department of Health, Education, and Welfare, Office of Education, under the provisions of the Cooperative Research Program. A Hogg Foundation grant supported essential preparations prior to the start of the second year's work. Preliminary drafts of this article were written during the author's USOE-sponsored postdoctoral research training fellowship at the University of Oregon. Completion of the work was partially supported by a grant from the University Research Committee of the Graduate School, University of Wisconsin, Madison.

role of contact with pupils rather than mere forgetting in this decline, one arrives at the following questions: (*a*) What is the direction of influence—from teachers to pupils or from pupils to teachers? (*b*) Is influence from pupils to teachers found more often in the classrooms of beginning teachers than in those of experienced teachers? [2] Finally, since teacher warmth may be especially significant for lower-class pupils (Gage, 1965), (*c*) does the direction of influence differ according to the pupils' social-class background?

To answer these questions, a 2-year study was conducted, starting in 1964 (Yee, 1966). Preliminary analyses indicated the need to distinguish between the direction as well as the source of interpersonal influence. Thus, for the direction of influence, we consider influence to be congruent when attitudes of source and influence become more positively correlated from pre- to posttest occasion and incongruent when attitudes become less positively correlated over time. With these distinctions, the following hypotheses were tested for total sample, three levels of teachers' years of teaching experience, and two levels of pupils' social class in one- and two-factor analyses.

Hypotheses

H_1 : Teacher-class pairs showing teacher influence toward either congruity (TC) or incongruity (TI) are more frequent than those showing pupil influence toward either congruity (PC) or incongruity (PI), that is, TC + TI > PC + PI.

H_2 : TC > PC.

H_3 : TI > PI.

METHOD

Instruments

Teachers' attitudes were measured with (*a*) the MTAI and (*b*) a semantic differential (Osgood, Suci, & Tannenbaum, 1957) prepared for this study with My Class as the concept and 17 bipolar adjectives highly loaded on the evaluative dimension.

Pupils' attitudes were measured with the 100-item About My Teacher (AMT) inventory (Beck, 1964) developed on five dimensions of teacher's merit: affective, cognitive, disciplinary, innovative, and motivational. This inventory yielded a total score (P_0, P_0'; unprimed symbols indicate pretest and primed symbols indicate posttests) and 11 subscores obtained on the basis of multiple-factor analyses (principal axis, rotated by Kaiser's Varimax method[3]) of the mean pupil ratings of their teachers. Identical or very similar factors were extracted from separate analyses of the middle-class and lower-class pupils' responses.

The factor analyses provided the basis for 11 measures of dimensions of pupils' perceptions of their teachers:

P_1 : popularity and effectiveness in instruction (17 items, such as "Do you think your teacher understands people your age? Does your teacher make sure everybody understands the teacher?")

P_2 : personal popularity and warmth toward children (6 items, such as "Does your teacher seem to like children?")

P_3 : irritability and moodiness (3 items, such as "Is your teacher often cross?")

P_4 : explaining ability and communication, as measured with negatively stated items (9 items, such as "When you ask your teacher a question, do you often just get more confused?")

P_5 : explaining ability and communication, as measured with positively stated items (8 items, such as "Does your teacher make difficult things easy to understand?")

P_6 : effectiveness in developing an atmosphere of responsible pupil conduct, or pupils' perceptions of their own orderliness (9 items, such as "Do the children behave well for your teacher?")

P_7 : disciplining behavior (3 items, such as "Does your teacher succeed in keeping the pupils under control?")

P_8 : innovative behavior, tendency to use audiovisual materials and field trips (3 items, such as "Does your class go on field trips that help you understand what you are studying?")

P_9 : innovative behavior, tendency to individualize instruction in the choice of materials and methods (3 items, such as "Do all the pupils in the class use the same book at the same time?")

P_{10} : motivating pupils' interest and enthusiasm, as measured by positively stated items (6 items, such as "Does your teacher make you feel like working real hard at your school work?")

P_{11} : motivating pupils' interest and enthusiasm, as measured with negatively stated items (3 items, such as "Is your teacher making school work less interesting for you this year?")

For measures of teachers' attitudes, the total MTAI scores were supplemented by three MTAI factors extracted by Horn and Morrison (1965): Factor I, Traditionalistic Versus Modern Beliefs about Child Control (T_1 , T_1'); Factor II, Un-

[2] For suggesting teacher experience as a factor in our investigation, the author is indebted to R. L. Debus, University of Sydney.

[3] With computer program by D. J. Veldman on file at the Computation Center, The University of Texas.

favorable Versus Favorable Opinions about Children (T_2, T_2'); and Factor III, Punitive Intolerance Versus Permissive Tolerance for Child Misbehavior (T_3, T_3').

Sample

Data for this study were secured from 102 teachers and their pupils in 32 schools of middle-class (MC) neighborhoods located in the San Francisco bay area and central Texas (in Grade 4, $n = 33$; Grade 5, $n = 36$; Grade 6, $n = 33$) and 110 teachers and their pupils in 18 schools of lower-class (LC) neighborhoods located in central Texas (in Grade 4, $n = 39$; Grade 5, $n = 38$; Grade 6, $n = 31$; Grade 7, $n = 2$). Recruiting procedures conducted with school administrators created no known systematic bias in the sample selected. The sample obtained is believed to be a fair representation of LC and MC pupils and their teachers, even though rigorous random selection was not possible.

Social-class status was determined by consultation with school administrators and informal inspection of neighborhoods. Family income ($4,000 or less annually for lower class; $6,000 or more for middle class) and father's occupation (blue collar and unskilled for lower class; white collar and professional for middle class), as ascertained from school administrators, were the main criteria for establishing social-class status.

For analyses to be discussed, subsamples were classified by pupils' social-class background and teachers' years of experience, as follows: for LC, 0–1 years (five with 1 year's experience), $n = 25$; 2–8 years (average of 4.6 years), $n = 36$; 9+ years (average of 19.9 years), $n = 49$; and for MC, 0–1 years (four with 1 year's experience), $n = 39$; 2–8 years (average of 5.5 years), $n = 31$; 9+ years (average of 17.2 years), $n = 32$.

Procedure

The frequency-of-change-in-product-moment (FCP) technique was developed to tabulate each teacher-class unit under one form of teacher or pupil influence, that is, TC, TI, PC, or PI. This technique and others are more fully discussed in Yee and Gage (1968).

The following procedures were originated to derive frequencies for chi-square tests of significance:

1. Raw scores of teachers' and pupils' attitudes (class means) were converted to standard scores.

2. The nature of or *direction of influence*—congruent or incongruent—was determined by seeing if cross-products of posttest z scores were more positive or negative than cross-products of pretest z scores. If the cross-product of posttest z's, $z_{T_n'}z_{P_n'}$, was more positive than $z_{T_n}z_{P_n}$, the direction of influence was said to be congruent, that is, the relationship between the teacher and her class helped make the overall correlation more positive. If the cross-product of posttest z's was more negative, the direction of influence was called incongruent, that is, the relationship between the teacher and her class helped make the overall correlation more negative. This manner of assessing direction of influence is logically connected with the basic formula for product-moment correlation coefficients, that is, $r = \Sigma z_x z_y / N - 1$.

3. The *source of influence* was determined by taking cross-lagged z products, $z_{T_n}z_{P_n'}$ and $z_{P_n}z_{T_n'}$. When direction of influence was congruent, the more positive product was classed as source, that is, it helps to increase the cross-lagged correlation where effector's z score is from pretest occasion and z score of party influenced is posttest. When direction of influence was incongruent, the more negative product was classed as source, that is, it helps to increase the cross-lagged correlation where effector's z score is from posttest occasion and z score of the one influenced is pretest.

The three hypotheses were tested with the general computing formula for chi square (Guilford, 1965, p. 230). The hypotheses as stated call for a directional or a one-tailed test of significance; therefore, the .05 level of significance requires a chi-square value of at least 2.71 with 1 *df*. In computing a chi-square value, Yates' correction for continuity (Guilford, 1965, pp. 237–239) was applied to the frequencies.

Results and Discussion[4]

Reliability, Rectilinearity, and Stability

Coefficients of internal consistency for the responses to both teacher inventories were computed with the Guttman (1945) L_4 formula. The coefficients for the My Class inventory and the MTAI were all in the high .80's.

Coefficients of pupil agreement in rating their teachers on the AMT were computed with Horst's (1949) formula. The Horst r for the total score (P_0) obtained at pretest was .86 for the 110 classes of LC pupils and .89 for the 102 classes of MC pupils; for total scores at posttest, the respective r's were .88 and .91. When separate Horst r's were computed for the various factor scores, these r's averaged .82 for the 110 teachers of LC pupils and .85 for the 102 teachers of MC pupils.

The rectilinearity of the various relationships between teachers' attitudes and pupils' perceptions was estimated by inspection of over 50 scatter-plots machine drawn for a

---

[4] H. Albert Napier and W. E. Geeslin provided data processing and computer assistance for this study with the help of Janice Willenborg.

random sample of r's. No curvilinear relationship was found.

Stability coefficients were higher for the measures of teachers than those of pupils. For the total sample ($N = 212$), $r_{T_0T_0'} =$ .79 and $r_{P_0P_0'} = .68$. As expected, teachers' measures increased in stability with teaching experience. For teachers of 0–1 years' experience ($n = 65$), 2–8 years ($n = 67$), and 9–46 years ($n = 81$), $r_{T_0T_0'}$ equalled .71, .81, and .84, respectively. This regular increase in MTAI stability with increasing teacher experience is also evident when the teachers in each social-class group are divided according to experience. MC pupils' attitudes tend to be more stable than those of LC pupils, for example, for LC, $r_{P_0P_0'} =$ .62 and for MC, .71.

FCP Analyses

The direction of attitude change when found in statistically significant results is in general from teachers to pupils, that is, teachers' attitudes cause pupils' attitudes to change more than pupils' attitudes cause teachers' attitudes to change. The most striking FCP results were obtained in analyses of attitude relationships with teachers' T_0 and T_1 scores.

In general, the attitude means for teachers of LC pupils were significantly lower than those for teachers of MC pupils (Yee, 1968). For example, respective T_1 means for LC and MC pupils' teachers with 9+ years' experience were as follows: $\bar{T}_1' = .45$, $SD = 9.80$; $\bar{T}_1' = 6.41$, $SD = 9.12$. The low T_1 scores for teachers of LC pupils indicate such teachers possess traditionalistic and inflexibly negative attitudes toward child control. The higher T_1 scores for MC pupils' teachers show a more permissive, positive, and flexible attitude toward controlling children. Thus, analyses with this teacher attitude indicate great contrast between LC and MC subsamples and many instances of teacher dominance over LC pupils in attitude relationships.

Results of analyses with other teacher measures indicate the direction of influence to be from teachers to pupils in the preponderance of frequencies favoring teacher influence, but fewer statistically significant results were found with them than were found with T_0 and T_1. Since T_1 was the primary factor extracted by Horn and

TABLE 1

CORRELATIONS AND FREQUENCIES-OF-CHANGE-IN-PRODUCT-MOMENT RESULTS FOR TEACHER VARIABLE (T_1) AND PUPIL VARIABLE (P_5)

Pupils' social class	Teachers' years of experience	Correlations					Frequencies				Chi squares for hypotheses[a]		
		N	$r_{T_1T_1'}$	$r_{P_5P_5'}$	$r_{T_1P_5}$	$r_{T_1'P_5'}$	TC	TI	PC	PI	H_1	H_2	H_3
Both	0–46	212	.79	.59	.00	.06	60	65	50	37	*6.46*	.74	*7.15*
LC	0–46	110	.82	.55	.14	.16	37	38	23	12	*13.33*	*2.82*	*12.50*
MC	0–41	102	.74	.64	−.12	−.05	25	26	30	21	.01	.29	.34
Both	0–1	64	.71	.68	−.14	.02	21	15	14	14	.77	1.03	.03
Both	2–8	67	.75	.53	.09	.25	20	21	14	12	*2.93*	.74	1.94
Both	9–46	81	.85	.55	−.03	−.05	24	33	15	9	*12.64*	1.64	*12.60*
LC	0–1	25	.84	.67	.10	.26	8	7	6	4	.64	.07	.36
LC	2–8	36	.77	.57	.26	.37	18	10	3	5	*10.03*	*9.33*	1.07
LC	9–46	49	.83	.50	.01	−.07	14	18	11	6	*4.00*	.16	*5.04*
MC	0–1	39	.59	.72	−.24	−.15	11	9	9	10	.03	.05	.05
MC	2–8	31	.75	.58	−.08	.05	4	11	12	4	.03	*3.06*	2.40
MC	9–41	32	.86	.64	−.08	.01	7	10	8	7	.03	.07	.24

Note.—Abbreviated: LC = lower class, MC = middle class, TC = teacher congruity, TI = teacher incongruity, PC = pupil congruity, PI = pupil incongruity, H_1 = Hypothesis 1, H_2 = Hypothesis 2, H_3 = Hypothesis 3.

[a] H_1 : TC + TI > PC + PI; H_2 : TC > PC; and H_3 : TI > PI. Yates' correction applied to chi squares; chi square equals 2.71 at .05 level of significance, 3.84 at .02 level, 5.41 at .01 level, and 6.64 at .001 level, one-tailed with 1 df. Chi squares with $p < .05$ level are in italics.

Morrison (1965) and closely resembles the primary factor extracted by this writer in a preliminary factor analysis of the MTAI, T_1 accounts for more variance between teachers and appears to provide more salient response consistencies toward children than other teacher measures.

Table 1 illustrates the tabulation of frequencies by the FCP technique.

In Table 1, significant results show teacher influence represented by T_1 causing pupils' perceptions of teachers' explaining ability (P_5) to shift in congruent and incongruent directions. For the total sample, teachers cause pupils to shift significantly toward incongruity. Across subsamples, there is a definite pattern in the direction of frequencies favoring teacher influence. With LC pupils, teachers' influence predominates over pupils in both congruent and incongruent directions. Results of analyses with MC pupils indicate little one-sided dominance by either teachers or pupils.

One of the study's few instances in which statistically significant results favor pupil influence can be seen for the MC, 2–8 group in Table 1 under H_2.

Subsamples by Pupils' Social Class

Table 2 presents results for the subsample with LC pupils and their teachers. The asterisks indicate the statistically significant chi-square results found. Analyses with T_0 and T_1 show strong teacher dominance over LC pupils, especially in incongruent influence. Such results contrast sharply with results from similar analyses for MC pupils, as shown in Table 3. Only in analyses with MC pupils' attitudes toward teachers' innovative (P_8 and P_9) and motivational (P_{10}) merit do predominant results in Table 3 favor teacher influence. Since innovative and motivational pupil attitudes are based on perceptions of teacher behavior most likely to change over time (see earlier description of these pupil factors), such results are less significant than if comparable results with other factors had been found.

The great contrast between results in Tables 2 and 3 indicate considerable differences in teacher-pupil interaction, which are reflected in the contrasting attitude means where attitudes of the LC subsample are generally lower than those found for the MC subsample (Yee, 1968).

TABLE 2

Summary of Frequencies-of-Change-in-Product-Moment Results for the Subsample of 110 Teacher-Class Units with Lower-Class Pupils

Pupil variables	Hypothesis 1					Hypothesis 2					Hypothesis 3				
	Teacher Variables														
	T_0	T_1	T_2	T_3	C	T_0	T_1	T_2	T_3	C	T_0	T_1	T_2	T_3	C
P_0	****	*****	—	—	—	—	—	—	—	—	****	*****	—	—	—
P_1	—	**	—	—	—	—	—	—	—	—	—	***	—	—	—
P_2	—	****	—	—	—	—	—	—	—	—	—	**	—	—	—
P_3	****	*****	—	—	—	*	—	—	—	—	**	*****	—	—	—
P_4	*	*****	—	**	—	—	****	—	**	—	—	****	—	—	—
P_5	**	*****	**	****	—	—	*	—	—	—	—	*****	—	**	—
P_6	—	*****	—	—	—	—	***	—	—	—	*	**	—	—	—
P_7	*****	*****	*	****	—	*	****	—	*	—	*****	*****	***	**	—
P_8	***	*****	—	—	—	—	—	—	—	—	****	*****	*	—	***
P_9	****	*****	—	****	—	—	—	—	—	—	****	*****	—	****	—
P_{10}	*****	*****	—	***	—	****	****	—	*	—	****	*****	—	—	—
P_{11}	**	*****	—	*	—	—	****	—	—	—	—	****	—	—	—

Note.—Asterisks indicate significance level of chi-square results favoring teacher influence.
* $p < .05$.
** $p < .02$.
*** $p < .01$.
**** $p < .001$.
***** $p < .0001$.

TABLE 3

SUMMARY OF FREQUENCIES-OF-CHANGE-IN-PRODUCT-MOMENT RESULTS FOR THE SUBSAMPLE OF 102 TEACHER-CLASS UNITS WITH MIDDLE-CLASS PUPILS

Pupil variables	Hypothesis 1					Hypothesis 2					Hypothesis 3				
	Teacher variables														
	T_0	T_1	T_2	T_3	C	T_0	T_1	T_2	T_3	C	T_0	T_1	T_2	T_3	C
P_1	*	—	—	—	—	—	—	—	—	—	—	—	—	—	—
P_6	—	—	—	—	—	*	—	**	—	—	—	—	—	—	—
P_8	****	*	—	**	*	****	*	—	—	—	—	—	—	—	—
P_9	*****	**	**	**	****	****	**	***	**	**	*	—	—	—	*
P_{10}	***	**	***	*	—	**	—	—	—	—	—	***	****	**	—
P_{11}	*	—	—	—	—	—	—	—	—	—	—	—	—	—	—

Note.—Asterisks indicate significance level of chi-square results favoring teacher influence; no significant results found for analyses with P_0, P_2, P_3, P_4, P_5, and P_7.

* $p < .05$.
** $p < .02$.
*** $p < .01$.
**** $p < .001$.
***** $p < .0001$.

Subsamples by Pupil's Social Class and Teacher's Experience

Beginning teachers. LC pupils taught by novice teachers are influenced more than MC pupils. In T_0 and T_1 analyses for the subsample, five significant chi squares supporting H_1 were found, two for H_2, and seven for H_3. In analyses with the same attitudes, only one significant result was found for the relationships between novice teachers and MC pupils; and that was for H_1 with P_9. No results favored LC or MC pupils' influence.

The causal dominance of teachers over LC pupils may be more understandable with closer examination of results for the subsample with LC pupils and beginning teachers. It was found that LC pupils' disciplinary (P_7) and motivational (P_{10}) attitudes shifted in the congruent direction (that is, became more positively correlated to teacher's attitudes), but their affective (P_1 and P_2), cognitive (P_4 and P_5), and innovative (P_9) attitudes shifted in the incongruent direction (that is, became less positively correlated to teacher's attitudes). Such results indicate that from pre- to posttest occasion, interaction became more teacher dominated, pupils became more conforming, and classroom climate grew colder. School became less appealing for the LC student who in acquiescing to the negativity of their teacher's attitudes as indicated by pretest scores may have contributed to the slight improvement of their teacher's attitudes as indicated by posttest scores (e.g., $\overline{T}_1 = 4.68$, $\overline{T}_1' = 5.28$).

Experienced teachers. Results for four subsamples with experienced teachers, that is teachers with 2–8 years' experience and those with 9+ years' experience, also indicated differences between teacher-pupil interaction for LC and MC pupils. The magnitude and extent of teacher dominance over LC pupils in T_1 relationships become greater as teachers' experience increases.

In analyses with T_1 scores of teachers with 2–8 years' experience and all LC pupils' scores, seven significant chi squares supporting H_1 were found: five for H_2, and six for H_3. For the counterpart MC group, not one result favored teacher influence. However, three results just barely significant at the .05 level favored pupil influence.

The contrast between teacher-pupil interaction for LC and MC pupils becomes even greater with teachers of 9+ years' experience. Nine of the 12 FCP results with senior teachers' T_1 measures and LC pupils' measures supported H_1 and 9 supported

H_3. No significant results for H_2 or in support of pupil influence were found. The predominance of significant results for H_3 indicates considerable incongruent teacher-pupil interaction operating between senior teachers and LC pupils. Results for MC pupils and their senior teachers show only one significant chi square, and that is for H_1 with pupils' P_9 attitude.

Discussion and Conclusions

When a definite direction of influence in teacher-pupil interaction was found in FCP analyses, teachers influence their pupils much more in schools located in LC neighborhoods than those located in MC neighborhoods. In schools for LC pupils, teachers' less positive attitudes of warmth, permissiveness, and favorability toward pupils tended to make pupils' attitudes toward their teacher become more unfavorable. In schools for MC pupils, the teachers' more positive attitudes made less difference, that is, had less effect on pupils' attitudes.

The factor of teachers' experience accounted for far less of the variance between teachers and classroom interaction than pupils' social class. Between the two levels of pupils' social class, attitude relationships of all three levels of teacher experience contrasted greatly; but within the same level of pupils' social class, levels of teacher experience resembled each other. Teacher dominance over LC pupils appears to become more pronounced as teacher experience increases. The more negative attitudes of this study's teachers with 9+ years' experience working with LC pupils and their incongruent attitude relationships with LC pupils raise serious questions concerning such teachers' placement with LC pupils and value as teachers in general.

Perhaps consideration of the pupils' characteristics helps explain these results in part (Riessman, 1962). It may be argued that LC pupils have less potent sources of adult warmth and support at home and hence depend more on, and are influenced by, such adult influence at school. The more vulnerable self-concept, or weaker ego of the LC pupil makes him more open to his teacher's influence as a determiner of his attitude toward his teacher. The better established orientation of the MC child toward adults in general, both parents and teachers, makes his attitudes toward his teacher more stable and less susceptible to the influence of the particular teacher he happens to have in any given year.

Nevertheless, the more significant factors determining such differences in teacher-pupil interaction appears to be in the area of teachers' characteristics rather than the area of pupils' characteristics. The relative differences between the pretest attitudes of MC and LC pupils' teachers are greater than the differences between the pretest attitudes of MC and LC pupils. One explanation for such differences between teachers may be what administrators and supervisors emphasize in evaluating teacher performance in systems with MC pupils and in those with LC pupils. Turner (1965) suggested that the greater need to emphasize and help overcome pupils' deficiencies in skill subjects as seen by administrators and supervisors in schools with working-class children creates a "criterion space" less concerned with teachers' affective merit. Kliebard's (1967) critical review of curriculum strategies for disadvantaged learners lends support to Turner's argument. In schools serving MC families, pupils' deficiencies in skills and intellectual background are not major problems, so greater emphasis is given the personal-social characteristics of teachers.

Although much has been said and written concerning the pre- and inservice preparation of teachers for work with disadvantaged children (e.g., Riessman, 1967), such preparation may be for naught unless school administrators develop pedagogical and employment policies that recognize the affective needs of disadvantaged pupils as well as their cognitive needs. The practical significance of these findings and interpretations is that the teacher's attitudes of warmth and permissiveness are even more important to LC children that to MC children. Zigler and Kanzen (1962) found a significant interaction between the type of reinforcer used and the social class of *S*. The praise reinforcers, such as "good" and "fine," were more reinforcing than the

correct reinforcers, such as "correct" and "right," with LC children, while the correct reinforcers were more effective than the praise reinforcers with MC children.

The hypothesis that LC pupils are discriminated against with respect to the teachers and classroom environments assigned them appears to be supported by this study's results. Insofar as such teacher attitudes can be brought into the classroom through selection and training procedures, the effort should especially be made to place the "better" teachers in schools located in LC neighborhoods.

REFERENCES

BEAMER, G. C., & LEDBETTER, E. W. The relation between teacher attitudes and the social service interest. *Journal of Educational Research*, 1957, **50**, 655-666.

BECK, W. H. Pupils' perceptions of teacher merit: A factor analysis of five hypothesized dimensions. Unpublished doctoral dissertation, Stanford University, 1964, *Dissertation Abstracts*, 1965, 25, 5668.

BUSH, R. N. *The teacher-pupil relationship*. New York: Prentice-Hall, 1954.

DAY, H. P. Attitude changes of beginning teachers after initial teaching experience. *Journal of Teacher Education*, 1959, **10**, 326-328.

DELLA PIANA, G. M., & GAGE, N. L. Pupils' values and the validity of the Minnesota Teacher Attitude Inventory. *Journal of Educational Psychology*, 1955, **46**, 167-178.

FLANDERS, N. A. *Teacher influence, pupil attitudes, and achievement*. Document No. OE-25040, 1965, Government Printing Office, Washington, D. C.

GAGE, N. L. Psychological research on teacher education for the great cities. *Urban Education*, 1965, **1**, 175-196.

GAGE, N. L., RUNKEL, P. J., & CHATTERJEE, B. B. Changing teacher behavior through feedback from pupils: An application of equilibrium theory. In W. W. Charters, Jr., & N. L. Gage (Eds.), *Readings in the social psychology of education*. Boston: Allyn & Bacon, 1963.

GETZELS, J. W., & JACKSON, P. W. The teacher's personality and characteristics. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally, 1963.

GUILFORD, J. P. *Fundamental statistics in psychology and education*. (4th ed.), New York: McGraw-Hill, 1965.

GUTTMAN, L. A basis for analyzing test-retest reliability. *Psychometrika*, 1945, **10**, 255-282.

HEIL, L. M., & WASHBURNE, C. Brooklyn College research in teacher effectiveness, *Journal of Teacher Education*, 1962, **55**, 347-351.

HORN, J. L., & MORRISON, W. E. Dimensions of teacher attitudes. *Journal of Educational Psychology*, 1965, **56**, 118-125.

HORST, P. A generalized expression of the reliability of measures. *Psychometrika*, 1949, **14**, 21-32.

HOYT, C. J., & COOK, W. W. The stability of MTAI scores during two to seven years of teaching. *Journal of Teacher Education*, 1960, **11**, 487-491.

KLIEBARD, H. M. Curriculum differentiation for the disadvantaged. *Educational Forum*, 1967, **31**, 47-54.

OSGOOD, C. F., SUCI, G. J., & TANNENBAUM, P. H. *The measurement of meaning*. Urbana: University of Illinois Press, 1957.

RABINOWTIZ, W., & ROSENBAUM, I. Teaching experience and teachers' attitudes. *Elementary School Journal*, 1960, **60**, 313-319.

RIESSMAN, F. Teachers of the poor: A five-point plan. *Journal of Teacher Education*, 1967, **18**, 326-336.

RIESSMAN, F., *The culturally deprived child*. New York: Harper, 1962.

RYANS, D. G. *Characteristics of teachers*. Washington, D. C.: American Council on Education 1960.

SMITH, B. O. A concept of teaching. *Teachers College Record*, 1960, **61**, 229-241.

TURNER, R. L. Characteristics of beginning teachers: Their differential linkage with school-system types. *School Review*, 1965, **73**, 48-58.

WHITE, R. K., & LIPPITT, R. *Autocracy and democracy: An experimental inquiry*. New York: Harper, 1960.

WITHALL, J., & LEWIS, W. W. Social interaction in the classroom. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally, 1963.

YEE, A. H. Factors involved in determining the relationship between teachers' and pupils' attitudes. U. S. Department of Health, Education, and Welfare, Office of Education Research Project No. 5-8346, Austin: The University of Texas, 1966.

YEE, A. H. Interpersonal attitudes of teachers and advantaged and disadvantaged pupils. *Journal of Human Resources*, 1968, in press.

YEE, A. H., & GAGE, N. L. Techniques for estimating the source and direction of causal influence in panel data. *Psychological Bulletin*, 1968, in press.

ZIGLER, E., & KANZER, P. The effectiveness of two classes of verbal reinforcers on the performance of middle- and lower-class children. *Journal of Personality*, 1962, **30**, 156-163.

(Received July 20, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 4, 283–289

"WHAT IS LEARNED" IN MATHEMATICAL DISCOVERY[1]

WILLIAM G. ROUGHEAD AND JOSEPH M. SCANDURA
North Georgia College *University of Pennsylvania*

2 questions were asked: (*a*) can "what is learned" in mathematical discovery be identified and taught by exposition with equivalent results, (*b*) how does "what is learned" depend on prior learning and on the nature of discovery? The major hypothesis was that discovery *S*s may discover derivation rules for deriving classes of solutions but only when the solutions are not initially known. 4 programs, (specific) rule-given (R), discovery (D), guided discovery (G), and exposition of derivation rule (E) were administered to 7 groups. 1 group received program R alone; the others received R with 1 of the other programs. Both orders of presentation were represented: RD, DR; RG, GR; RE, ER. All Ss were required to derive new solutions within the scope of the derivation rule. As hypothesized, Groups R and RD performed at 1 level which was reliably ($p < .001$) below the common level of the other 5 groups. Theoretical and practical implications were discussed.

One of the fundamental assumptions underlying many of the new mathematics curricula is that discovery methods of teaching and learning increase the students' ability to learn new content (e.g., Beberman, 1958; Davis, 1960; Peak, 1963). The last decade of research on discovery learning, however, has produced only partial and tentative support for this contention. Even where the experiments have been relatively free of methodological defects, the results have often been inconsistent (e.g., see Ausubel, 1961; Kersh & Wittrock, 1962). More particularly, the interpretation of research on discovery learning has been made difficult by differences in terminology, the tendency to compare identical groups on a variety of dependent measures, and vagueness as to what is being taught and discovered.

While most discrepancies due to differences in terminology can be reconciled by a careful analysis of what was actually done in the experiments (e.g., Kersh & Wittrock, 1962) and thus present a relatively minor problem, the failure to equate original learning has often made it difficult to interpret transfer (and retention) results in an unambiguous manner. Thus, several studies (e.g., Craig, 1956; Wittrock, 1963) have shown that rule-given groups perform better on "near" transfer tests than do discovery groups. The obtained differences, however, may have been due to the fact that the discovery groups did not learn the originally presented materials as well as the rule-given groups.

When the degree of original learning was equated, Gagné and Brown (1961) found that their discovery groups were better able to derive new formulas than were their rule- (i.e., formula) given groups. They attributed this result to differences in "what was learned" but added that they were unable to specify precisely what these differences were. On the basis of an analysis of the experimental programs used by Gagné and Brown (1961), Eldredge (1965) hypothesized that the differences found by

[1] This paper is based largely on a PhD dissertation submitted to the Florida State University by the first author under the chairmanship of the second author. The first author was primarily responsible for the conduct of the experiment and the analysis of the data. The second author was primarily responsible for formulating the problem and for the preparation of this report. This research was supported, in part, by a United States Office of Education grant to the second author.

The authors are indebted to Robert M. Gagné and Gabriel Della-Piana for making copies of their experimental materials available. The authors would also like to thank Joan Barksdale and John Durnin for their general assistance and Joanna P. Williams for reading a draft of this paper. The participation of Barksdale and Durnin was made possible by a United States Office of Education Graduate Training grant to the University of Pennsylvania in Mathematics Education Research.

Gagné and Brown (1961) were due to uncontrolled factors. Eldredge conjectured that if the treatment differences were limited to the order of presentation of the discovery hints and the to-be-learned formulas, no differences in transfer ability would result. However, Eldredge's results contradicted his hypothesis. In subsequent studies, Gutherie (1967) and Worthen (1967) obtained similar sequence effects.

Using the Set-Function Language (SFL) characterization of a rule as a guide, Scandura (1966) proposed an analysis of discovery learning that seems to be in accord with experimental findings.

In the SFL, the rule is viewed as the basic unit of behavior; associations and concepts are shown to be special cases (of the rule). The denotation of a rule is defined as a set of *functionally distinct* stimulus-response pairs—the instances of the rule. The rule construct itself is characterized as an ordered triple (D,O,R) where D refers to the set of those stimulus properties which *determine* the corresponding responses, and O refers to the *operation* or transformation by which the derived stimulus properties or (internal) *responses* in the set R are derived from the properties in D (for more details, see Scandura, 1966, 1967a, 1968a, b, and c).

The main point of the analysis was that in order to succeed, discovery *S*s must learn to derive solutions (i.e., responses) whereas solution-given *S*s need not. In attaining criterion, discovery *S*s may discover a derivation rule by which solutions to new, though related, problems may be derived. Under these circumstances, discovery *S*s would be expected to perform better than expository *S*s on tasks which are within the scope of such a derivation rule. If the new problems presented have solutions beyond the scope of a discovered derivation rule, however, there would be no reason to expect discovery *S*s to have any special advantage.

This study was concerned with two major questions. First, can "what is learned" in mathematical discovery be identified and, if so, can it be taught by exposition with equivalent results? Second, how does "what is learned" depend on prior learning and on the nature of the discovery treatment itself?

The SFL was used as an aid in analyzing the guided discovery programs used by Gagné and Brown (1961) and Eldredge (1965) to determine "what was learned." As a result of this analysis, an expository statement of the derivation rule was devised. It was possible, in the manner described by Scandura, Woodward, and Lee (1967), to determine on an a priori basis which kinds of transfer item could be solved by using this derivation rule and which could not.

Assuming that transfer depends only on whether or not the derivation rule is learned, then the order in which the formulas (i.e., the solutions) and the derivation rule are presented should have no effect on transfer so long as *S* actually learns the derivation rule. If, on the other hand, a discovery program simply provides an *opportunity* to discover and does not guide the learner through the derivation procedure, sequence of presentation might have a large effect on transfer. That is, if a capable and motivated *S* is given appropriate hints, he might well succeed in discovering the appropriate formulas and in the process discover the derivation rule. It is not likely, however, that he would exert much effort when given an opportunity to discover a formula he already knows. Something analogous may well have been involved in the studies by Eldredge (1965), Gutherie (1967), and Worthen (1967).

In particular, the following hypotheses were made. First, what was learned by guided discovery in the Gagné and Brown (1961) study can be presented by exposition with equivalent results. Second, presentation order is critical when the hints provided during discovery are specific to the respective formulas sought rather than relevant to a general strategy (i.e., derivation rule). Third, presentation order is *not* critical when the program effectively forces *S* to learn the derivation rule, regardless of whether the learning takes place by exposition or by discovery.

METHOD

Materials[2]

There were seven treatments. Each consisted of a common introductory program followed by

[2] Copies of the experimental materials used are included in Roughead's (1966) dissertation and in Scandura's (1967b) final report.

various combinations of four basic instructional programs. The introductory program was designed to generally familiarize *S*s with number sequences and with terminology used in the four basic programs. In particular, four concepts were clarified: sequence; term value, T_n; term number, n; and sum of the first n terms of a sequence, $\sum^n$.

Each of the four basic instructional programs was based on the same three arithmetic series and their respective summing formulas: $1 + 3 + 5 + \cdots + (2n - 1) \to n^2$; $2 + 6 + 10 + \cdots + (4n - 2) \to 2n^2$; $1 + 5 + 9 + \cdots + (4n - 3) \to (2n - 1)n$. Following Gagné and Brown (1961), each series was presented as a three-row display—e.g.

Term number	n:	1	2	3	4	5 ···
Term value	T_n:	2	6	10	14	···
Sum	$\sum^n$:	2	8	18	32	···

The rule and example (R) program consisted of the three series displays together with the respective summing formulas. The presentation of each summing formula was followed by three application problems—e.g., find the sum of $2 + 6 + 10$ ($= 2 \cdot 3^2 = 18$). The *S* was also required to write out each formula in both words and symbols, but no rationale for the formula was provided. A test of the three training formulas was included at the end of the R program.

The other three basic programs included differing kinds of directions and/or hints as to how the summing formulas might be determined. The expository (E) and (highly) guided discovery (G) programs were based on a simplified variant of that derivation rule presumably learned by the guided discovery *S*s in the Gagné and Brown (1961) study. The identified rule can be stated,

> . . . formulas for $\sum^n$ may be written as the product of an expression involving n [i.e., $f(n)$] and n itself. The required expression in n can be obtained by constructing a three columned table showing: (1) the first few sums $\sum^n$, (2) the corresponding values of n, and (3) a column of numbers $f(n) = \sum^n/n$ which when multiplied by n yields the corresponding values of $\sum^n$. Next, determine the expression $f(n) = \sum^n/n$ by comparing the numbers in the columns labeled n and $\sum^n/n$ and uncovering the (linear) relationship between them. The required formula is simply $\sum^n = n \cdot f(n)$.

As an example, consider the display,

Term number	n:	1	2	3	4	5 ···
Term value	T_n:	2	6	10	14	···
Sum	$\sum^n$:	2	8	18	32	···

The three-columed table would look like,

$f(n)$	n	$\sum^n$
2	1	2
4	2	8
6	3	18
8	4	32
⋮	⋮	⋮
$2n$	n	$2n^2$

The emerging pattern is $f(n) = 2n$; so, $\sum^n = 2n \cdot n = 2n^2$.

The E program consisted of a simplified statement of the derivation rule as it applied to each of the three training series. To insure that *S* learned how to use the derivation rule, a vanishing procedure was used which ultimately required *S* to apply the procedure without any instructions. The G program paralleled the E program in all respects. The only difference was that the G program consisted of questions whereas the E program consisted of yoked direct statements, each followed by a parallel question or completion statement to see whether *S* had read the original statement correctly. For example, the E statements, "When $n = 3$, you can multiply 6 times n to get $\sum^3 = 18$. What times n gives $\sum^3 = 18$?" corresponded to the question, "When $n = 3$, what times n gives $\sum^3 = 18$?" which appeared in the G program. Since the degree of overt responding was held constant, the only difference between the E and G programs was whether the information was acquired by reception or by reacting to a question (i.e., by discovery). The discovery (D) program, on the other hand, simply provided *S* with an opportunity to discover the respective summing formulas. The *S* was guided by questions and hints which were specific to the formulas involved (e.g., "the formula has a 2 in it") rather than relevant to any general strategy or derivation rule. The questions and hints were interspersed with liberal amounts of encouragement (e.g., "Good try," "you can do it," etc.) to provide motivation.

There were two transfer tests. The *within-scope* transfer test consisted of two new series displays which could be solved by the identified derivation rule. These series and their respective summing formulas were $3 + 5 + 7 + \cdots + (2n + 1) \to (n + 2) \cdot n$ and $4 + 10 + 16 + \cdots + (6n - 2) \to (3n + 1)n$. The *extra-scope* transfer test involved the series, $2 + 4 + 8 + \cdots + 2^n \to (2T_n - 2) = (T_{n+1} - 2)$ and $1/2 + 1/6 + 1/12 + \cdots + [1/n(n + 1)] \to n/(n + 1) = n^2 T_n$, which, strictly speaking, were beyond the scope of the identified derivation rule. A series of hints, paralleling those used in the D program, were constructed to accompany each test series.

The introductory and treatment programs were mimeographed and stapled together into separate $5\frac{1}{2} \times 8\frac{1}{2}$ inch booklets. The four transfer series were presented on separate pages in a test booklet in the same three-row form used in the learning programs. The hints were put on 5×7 inch cards, bound by metal rings.

Subjects, Design, and Procedure

The naive *S*s were 105 (103 females) junior and senior elementary education majors, enrolled in required mathematics education courses at the Florida State University, who volunteered to participate in the experiment. The data of seven other *S*s were discarded because they failed to meet a major premise on which the hypotheses were based. That is, they were poorly motivated and/or made

a large number of errors on the treatment programs.

The experimental *S*s were randomly assigned to the seven treatment groups. In addition to the common introductory program, the R group received only the R program. The other six treatment groups received the R program together with one of the other three basic instructional programs. The RE, RG, and RD groups received the R program followed by the E, G, and D programs, respectively, while the ER, GR, and DR groups received these same respective programs in the reverse order.

The *S*s were scheduled to come to the experimental room in groups of four or less and were arranged at the ends of two tables which were partitioned to provide separate study carrels. A brief quiz was used to screen out any *S*s who were already familiar with number series and/or formulas for summing them. Then, they were told:

> This is an experiment in learning mathematics. You will be given two programmed booklets to study. You are expected to try to learn. You should work at a good pace, but read everything for understanding.... If you have an error, don't change your answer, but write the correct answer under your original answer. If you can not respond to a question within a minute or so, put an "X" in the blank and continue. You should, of course, look back at the question after finding the answer to be sure you understand....

The *S*s worked at their own rate. The two *E*s observed the progress closely, provided general assistance and encouragement where needed, and recorded the times taken on the introductory and treatment booklets.

As soon as all of the *S*s in the testing group had completed the treatment programs, they were told to review for a test. After 2 minutes, the booklets were collected and the tests and hint cards were presented. The *S*s were instructed:

> On this test you will be timed. You also will be provided with hints to aid you when necessary. The less time it takes you and the fewer hints you need on a given problem, the better your score. You will be asked to find the formula for four new problems on this test. On each problem, you will have 5 minutes to find the correct summing formula. You should show any necessary work in your booklet. When you get an answer, raise your right hand immediately. Like this! Try it!... I'll tell you whether you are correct or incorrect. If incorrect, continue searching for the answer. Be sure to show me your answer quickly so that you get the best possible time score.... When I tell you that the 5 minutes are up, if you have not found the formula, you may begin using hints. You may use as many of the hints as you wish, and when you wish, after the 5 minute period. But remember, the fewer hints you use, the better your score.

Before continuing on to the second problem, each *S* read all of the hint cards pertaining to the first problem. The four *S*s in each testing group began each problem at the same time. If an *S* solved a problem before the others, he was allowed to read the rest of the hints for that problem and, then, was required to wait for the others to finish. Before being released, *S*s were asked not to discuss particulars of the experiment with others who might participate.

Three indexes of performance on the transfer tasks were obtained: (*a*) time to solution, (*b*) number of hints prior to solution, and (*c*) a weighted score similar to that used by Gagné and Brown (1961). The weighted score was equal to the time to solution in minutes plus a penalty of 4, 7, 9, or 10 depending on whether *S* used 1, 2, 3, or 4 hints, respectively. Theoretically, a range of scores from 0 to 20 was possible on this measure. Standard analysis of variance procedures were used to analyze the data after Cochran's C test failed to detect heterogeneity of variance.

RESULTS

Treatment Programs

All treatment groups performed at essentially the same level on the introductory program, both in terms of time to completion ($F = 1.74$, $df = 6/98$, $p > .05$) and number of errors ($F = 1.35$, $df = 6/98$, $p > .05$). Since the number of frames varied among the treatment programs, no overall comparisons were warranted.

Performance on Learning and Transfer Tests

The results on the within-scope transfer test conformed to prediction. Irrespective of the transfer measure used, the group (R) given the formula program only and the group (RD) given the formula program followed by the opportunity-to-discover program performed at one level ($F < 1$, $df = 1/28$) while the other five groups performed at a common ($F < 1$, $df = 4/70$) and significantly higher level ($F_{\text{time}} = 32.66$, $df = 1/98$, $p < .001$; $F_{\text{hints}} = 54.52$, $df = 1/98$, $p < .001$; $F_{\text{weighted}} = 57.99$, $df = 1/98$, $p < .001$). In particular, only that sequence effect involving Groups RD and DR was significant ($p < .01$).

While there were no overall treatment differences on the extra-scope transfer test (maximum $F = 1.31$, $df = 6/98$, $p > .05$),

TABLE 1

SUMMARY OF MEANS AND STANDARD DEVIATIONS OF TIME AND ERRORS ON INTRODUCTORY AND TREATMENT PROGRAMS; ITEMS CORRECT ON THE TEST OF CRITERION LEARNING; AND TIME, HINTS AND WEIGHTED SCORES ON WITHIN- AND EXTRA-SCOPE TRANSFER ITEMS

Program or test	Condition													
	R		RD		DR		RG		GR		RE		ER	
	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*
Introductory														
Time	8.93	2.20	9.00	2.71	7.13	1.59	7.73	2.38	8.73	2.93	7.27	1.00	8.13	2.09
Errors	1.20	.83	.80	1.04	.60	.80	.60	.61	.60	.71	.47	.88	.60	1.02
Treatment														
Time	23.07	7.42	29.33	8.70	33.00	7.70	36.40	10.29	41.47	8.30	35.87	8.70	40.20	8.60
Errors	1.47	1.86	1.20	1.04	1.53	2.22	2.80	2.51	2.67	2.87	1.73	1.65	2.20	1.48
Learning test														
Items correct	5.53	1.09	5.47	1.06	5.47	1.15	5.67	1.01	5.53	.81	6.00	0.00	5.73	.68
Within-scope														
Time	6.48	1.37	5.98	1.26	4.36	1.78	4.00	1.37	4.21	2.00	3.97	1.63	3.61	1.68
Hints	2.00	.79	1.73	.76	.80	.91	.50	.69	.60	.86	.64	.65	.50	.66
Weighted	12.87	3.56	11.65	3.30	6.97	4.31	5.73	3.40	6.27	4.68	6.20	3.82	5.37	3.91
Extra-scope														
Time	6.82	1.69	5.91	1.48	5.51	1.71	5.70	1.20	5.51	2.07	5.88	1.35	5.76	1.72
Hints	1.84	1.00	1.60	.88	1.20	.85	1.24	.73	1.37	1.11	1.44	.73	1.40	.71
Weighted	12.22	4.09	11.01	3.77	9.60	3.83	9.63	3.12	9.71	4.87	10.14	2.93	10.03	3.10

Note.—Abbreviated: R = rule given, D = discovery, G = guided discovery, E = exposition of derivation rule.

the contrast between Groups R and RD and Groups DR, RG, GR, RE, and ER attained a borderline significance level ($F_{\text{time}} = 3.66$, $df = 1/98$, $.05 < p < .10$; $F_{\text{hints}} = 4.02$, $df = 1/98$, $p < .05$; $F_{\text{weighted}} = 4.61$, $df = 1/98$, $p < .05$).[3] There were, however, no reliable performance differences between Groups DR and RD ($F < 1$).

[3] In a study on rule generality, Scandura, Woodward, and Lee (1967) obtained a similar extra-scope transfer effect. While no extra-scope transfer was almost universally the case, one of the rules (i.e., 50×50) introduced was apparently generalized (to $n \times n$) and thereby provided an adequate basis for solving an extra-scope problem. A recent study by Scandura and Durnin (1968) has demonstrated that, indeed, the form of a rule statement is an important determiner of generalization.

While not sufficient as presented, the derivation rule statement, introduced in this study, could also be generalized. In particular, the first hint available on the third transfer problem provided a basis for making appropriate modifications. Similarly, although Problem 4 involved fractional term values, the summing formula could be obtained by a relatively simple extension of the derivation rule statement.

For these reasons and because the results on the extra-scope test were subject to possible transfer effects of testing on the within-scope test, caution is advised in interpreting the extra-scope results. The extra-scope test was originally included to obtain experimental hypotheses and not definitive information. These comments, however, in no way apply to the clear results on the within-scope test.

These transfer effects cannot be attributed to differences in original learning. A learning test embedded within the common R program, indicated that *S*s had well learned the appropriate summing formulas to the three training series before they took the transfer tests. The group means ranged from 5.5 to 6 with a possible maximum of 6 and, minimum of 0. The error rates on the treatment programs were similarly low with an average of between one and two errors per program.

DICUSSION AND IMPLICATIONS

Two points need to be emphasized. First, "what is learned" during guided discovery can at least sometimes be identified and taught by exposition—with equivalent results. While this conclusion may appear somewhat surprising at first glance, further reflection indicates that we have always known it to be at least partially true. As has been documented in the laboratory (e.g., Kersh, 1958) as well as by innumerable classroom teachers of mathematics, it is equally as possible to give *S*s rules for deriving answers as it is to have them derive (i.e., discover) the answers themselves. No one to our knowledge, however, had ever seriously considered identifying "what is learned" in

deriving rules (i.e., formulas) in addition to the rules themselves. In the present study, the authors were apparently successful in identifying a (derivation) rule for deriving a class of more specific rules. No differences in the ability to derive new (within-scope) formulas could be detected between those *S*s who discovered a derivation rule and those who were explicitly given one. What was *not* done in this study was to consider the possibility that the discovery *S*s may have acquired a still higher order ability—namely, an ability to derive derivation rules. In any case, there are undoubtedly a large number of situations where, because of the complexity of the situation, "what is learned" by discovery may be difficult, if not impossible, to identify. In these situations, there may be no real alternative to learning by discovery.

Nonetheless, the value to transfer ability of learning by discovery does not appear to exceed the value of learning by some forms of exposition. Before definitive predictions can be made, careful consideration must be given to "what is learned," the nature of the transfer items, and the relationships between them. As we identify just what it is that is learned by discovery in a greater variety of situations, we shall be in an increasingly better position to impart that same knowledge by exposition.

The second point to be emphasized concerns the sequence effect—if a person already knows the desired responses, then he is not likely to discover another rule by which such responses may be derived, even if he has all of the prerequisites and is given an opportunity to do so. The reverse order of presentation may enhance discovery without making it more difficult to learn more specific rules at a later time. In short, prior knowledge may actually interfere in a very substantial way with later opportunities for discovery. Nonetheless, there may be some advantages inherent in learning more specific rules. Although data are practically nonexistent on this point, it is quite possible that specific rules may result in shorter latencies.

Why and how sequence affects "what is learned" is still open to speculation (e.g., Gutherie, 1967; Yonge, 1966). Our interpretation is as follows: When *S* is presented with one or more stimuli and is required to produce responses (e.g., formulas or specific rules) he does not already know, he *necessarily* must first turn his attention to deriving a rule (or derivation rule) by which he can generate the appropriate responses. In the process, *S* may discover a derivation rule, which is adequate for deriving other responses in addition to the ones needed. The kind and amount of guidance given would presumably help to determine the precise nature of the derivation rule so acquired. On the other hand, if *S* already knows the responses (i.e., has previously mastered more specific rules or "associations" by which the responses can be derived), it is not likely that he will waste much time trying to find another way to derive them. Under these conditions, it would seem that the only way to get *S* to learn a more general rule would be to change the context. Presumably, the expository and guided discovery *S*s in this study learned the derivation rule because this appeared to be the desirable thing to do. The authors believe that any theory based on the rule construct will have to invoke some such mechanism to account for sequence effects (e.g., see Scandura, 1968 a, b, and c).

The obtained sequencing result may also have important practical implications, as will be attested to by any junior high school mathematics teacher who has attempted to teach the "meaning" underlying the various computational algorithms after the children have already learned to compute. The children must effectively say to themselves something like, "I already know how to get the answer. Why should I care why the procedure works?" This is not to say that meaning should be taught first simply out of some sort of dislike for rote learning—for certain purposes rote learning may be quite adequate and the most efficient procedure to follow. The important point is that learning such things as how to multiply, without knowing what

multiplication means, may actually make it more difficult to learn the underlying meaning later on.

REFERENCES

AUSUBEL, D. P. Learning by discovery: Rationale and mystique. *The Bulletin of the National Association of Secondary School Principals*, 1961, **45**, 18–58.

BEBERMAN, M. *An emerging program of secondary school mathematics*. Cambridge, Mass.: Harvard University Press, 1958.

CRAIG, R. C. Directed versus independent discovery of established relations. *Journal of Educational Psychology*, 1956, **47**, 223–234.

DAVIS, R. B. The "Madison Project" of Syracuse University. *The Mathematics Teacher*, 1960, **53**, 571–575.

ELDREDGE, G. M. Discovery versus expository sequencing in a programmed unit on summing number series. In G. M. Della-Piana, G. M. Eldredge, & B. R. Worthen (Eds.), *Sequence characteristics of text materials and transfer of learning*. Salt Lake City: Bureau of Education Research, 1965.

GAGNÉ, R. M., & BROWN, L. T. Some factors in the programming of conceptual learning. *Journal of Experimental Psychology*, 1961, **62**, 313–321.

GUTHERIE, J. T. Expository instruction versus a discovery method. *Journal of Educational Psychology*, 1967, **58**, 45–49.

KERSH, B. Y. The adequacy of "meaning" as an explanation for the superiority of learning by independent discovery. *Journal of Educational Psychology*, 1958, **49**, 282–292.

KERSH, B. Y., & WITTROCK, M. C. Learning by discovery: An interpretation of recent research. *Journal of Teacher Education*, 1962, **13**, 461–468.

PEAK, P. (Ed.) *An analysis of new mathematics programs*. Washington, D. C.: National Council of Teachers of Mathematics, 1963.

ROUGHEAD, W. G. A clarification of part of the discovery versus exposition discussion in mathematics. Unpublished doctoral dissertation, Florida State University, 1966.

SCANDURA, J. M. Precision in research on mathematics learning: The emerging field of Psycho-Mathematics. *Journal of Research in Science Teaching*, 1966, **4**, 253–274.

SCANDURA, J. M. The basic unit in meaningful learning—association or principle? *The School Review*, 1967, **75**, 329–341. (a)

SCANDURA, J. M. *Teaching college students how to learn mathematics* ("What is learned" in mathematical discovery.) Project 6-8798. Washington: United States Office of Education, 1967. (Mimeo) (b)

SCANDURA, J. M. Using the rule to formulate research on meaningful learning: I. A set-function language. *Acta Psychologica*, 1968, in press. (a)

SCANDURA, J. M. Using the rule to formulate research on meaningful learning: II. Empirical research. *Acta Psychologica*, 1968, in press. (b)

SCANDURA, J. M. Using the rule to formulate research on meaningful learning: III. Analyses and theoretical direction. *Acta Psychologica*, 1968, in press. (c)

SCANDURA, J. M., & DURNIN, J. Extra-scope transfer in learning mathematical rules. *Journal of Educational Psychology*, 1968, in press.

SCANDURA, J. M., WOODWARD, E., & LEE, F. Rule generality and consistency in mathematics learning. *American Educational Research Journal*, 1967, **4**, 303–319.

WITTROCK, M. C. Verbal stimuli in concept formation: Learning by discovery. *Journal of Educational Psychology*, 1963, **54**, 183–190.

WORTHEN, B. R. A comparison of discovery and expository sequencing in elementary mathematics instruction. In J. M. Scandura (Ed.), *Research in mathematics education*. Washington: National Council of Teachers of Mathematics, 1967.

YONGE, G. D. Structure of experience and functional fixedness. *Journal of Educational Psychology*, 1966, **57**, 115–120.

(Received July 28, 1967)

Journal of Educational Psychology
1968, Vol. 59 No. 4, 290–296

MASSED VERSUS DISTRIBUTED PRACTICE IN COMPUTERIZED SPELLING DRILLS[1]

ELIZABETH JANE FISHMAN, LEO KELLER,[2] AND RICHARD C. ATKINSON
Stanford University

Conditions of massed and distributed practice were studied using a within-Ss design in a situation involving computerized spelling drills. In the distributed condition, 2 sets of 3 words each were presented once every other day over a period of 6 days. The learning trials on 6 other sets of words were massed so that all of the trials for that set occurred on the same day. Ss were 29 5th graders. The probability of a correct response for words in the massed condition was higher than that for the distributed condition during the learning sessions, but on retention tests (given 10 and 20 days later) the words learned under distributed practice were better remembered. A mathematical model of the learning process is presented and shown to provide a fairly adequate account of the experimental data.

Computer-assisted instruction (CAI) refers to an instructional procedure which utilizes a computer to control part, or all, of the selection, sequencing, and evaluation of instructional materials. Over the last 4 years, the Institute for Mathematical Studies in the Social Sciences at Stanford University has been developing a CAI system for regular classroom usage (Atkinson, 1967). One mode of this development is referred to by Suppes (1966) as the "drill and practice systems." These systems are intended to supplement the instruction which occurs in the classroom. They are designed to improve—through practice—the skills and concepts which are introduced by the classroom teacher.

Currently, computer controlled drills are being given to approximately 1,800 students in six schools in five different communities. Some of the students have been receiving daily drills in arithmetic (Suppes, Jerman, & Groen, 1966) while others have been receiving drills in spelling. This study made use of the equipment and students in the school which has been involved in drill and practice in spelling.

In the study to be reported here, the presentation routine for each spelling word was the same: An audio system presented the words, the student typed the word, and the computer evaluated the student's answer. If the response was correct, the computer typed "...C..."; if incorrect, "...X...", followed by the correct spelling of the word. If the response was not given within a predetermined length of time, the message "...TU...", meaning "time is up," was printed. A flow chart summarizing this procedure is given in Figure 1.

These CAI drill and practice systems lend themselves nicely to the study of many experimental variables. One persistent problem in designing instructional systems is the specification of optimal procedures for presenting material. Indeed, the spacing of learning sessions has already received considerable experimental investigation, yet the question of optimal spacing has not been resolved. For example, assume that we have 6 days in which to teach a list of 24 spelling words, and that each daily session is arranged so that 24 presentations can be made. What practice schedule would produce the best results? One might select a different set of four words each day and on that day present each word six times. At the other extreme, one could present each of the 24 words once per day. In both schemes a given word would be presented for study on six different occasions, but in one condition all of the repetitions for a given word would occur on 1 day whereas in the other scheme they would be distributed over 6 days.

[1] This research was supported by an Office of Education contract, No. OE5-10-050, and by a computing grant from the United States Public Health Service, Contract No. MH 06154.

[2] Now at the University of California, Irvine.

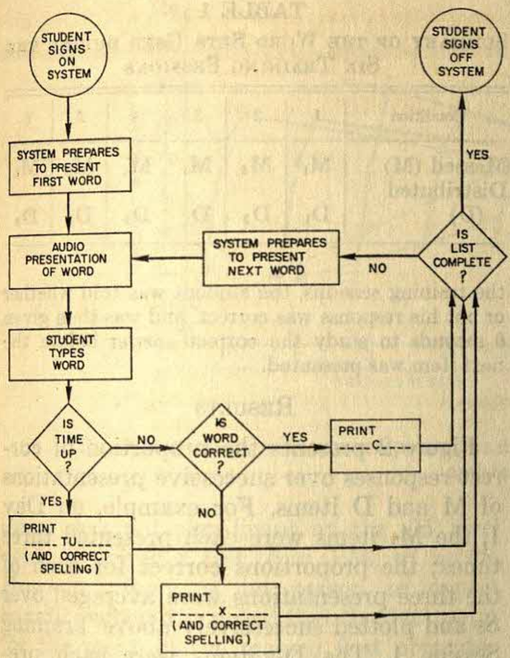


FIG. 1. Flow chart for presentation routine.

The two extremes could be called, respectively, massed and distributed practice, although this terminology is somewhat at variance with the classical usage of these terms. The preponderance of experimental evidence indicates that, for the same amount of practice, learning is better when practice is distributed rather than massed, although there are exceptions to the generalization. The purpose of the present study is to investigate this problem further and to evaluate optimum procedures for distributing instructional material in computer-based spelling drills.

METHOD

Subjects

The *Ss* were 29 students from a fifth-grade class in an East Palo Alto school. Approximately 50% of these students scored below grade level on standardized reading tests; 20% were reading at the second and third grade level.

The Computer System and Terminals

The computer which controlled the student terminals was a modified PDP-1 digital computer located at Stanford University. It was a time-sharing computer capable of handling over 30 different users simultaneously from a variety of input devices. The audio system for the spelling drills was controlled by a Westinghouse P-50 computer which, in turn, was linked to the PDP-1.

The four student terminals were located at an East Palo Alto school in a converted storeroom a short distance from the child's classroom. Each terminal consisted of a standard teletype machine and a set of earphones; both were linked to the computer at Stanford by telephone lines.

All four terminals were controlled by a single program on the PDP-1; each student user was serviced sequentially in a round-robin cycle. Due to the extremely rapid speed of the computer, the student received the impression that he was getting "full-time" service, although actually the computer was devoting only a small fraction of its running time to any one individual.

Daily Operation

A full-time monitor was on duty whenever the children were using the teletypes. Her presence was primarily a precautionary measure so that an adult would be available in case of an emergency. The actual check-in, presentation and evaluation of the drill, and the sign-out were all handled by the CAI system and occurred as follows.

The student entered the room, sat down at a free terminal, and put on his earphones. The machine printed out, "Please type your number." (This whole routine had been explained to the students during a 2-week orientatioin session.) After the student typed in his identification number and depressed the space bar—the latter operation was used as a termination signal for all student responses—the computer printed the student's name and the program was set in operation. The message, "If you hear the audio, please type an 'a' and a space," was then heard over the earphones. If the instructions were followed, the lesson began and each word was presented according to the sequence given in Figure 1.

The audio system presented a word, used the word in a sentence, and then repeated the word again. As soon as the audio was through, the machine typed a dash (—). This was the student's signal to begin his response. When he finished typing his answer, he depressed the space bar, and the computer evaluated the answer. A correct response was followed by the typed message, "...C...". An incorrect response was indicated by the message, "...X...," followed by several spaces and a correct spelling of the word. If a response was not given in 40 seconds, the message, "...TU..." was printed. As on an incorrect answer, this message was followed by several spaces and the correct spelling of the word. Following his response the student was given 6 seconds to study the correct answer before the next item was presented. Each time a new item was presented, all previous items were covered.

In the training sessions of this study, a "list" consisted of 12 such presentations; in the test sessions, 24 presentations. When the entire list had been presented, the machine printed out the following information for the student: his list number for the next session, the date and ending time,

and the number of words he spelled correctly on the day's session. The drills were collected by the monitor and at no time was the student given a copy of the words to study on his own.

Words

The words used in the experiment were taken from the New Iowa Spelling Scale (Greene, 1954). This scale is the product of the testing of some 238,000 pupils throughout the country in the early 1950s to determine the percentage of students that could spell a word correctly at each grade level. A list of the actual words used in the experiment can be found elsewhere (Fishman, 1967).

Experimental Design

The experiment involved a within-*Ss* design, (i.e., each *S* participated in all conditions). The two main conditions were those of massed (M) and distributed (D) practice. There were eight sets of words: six of them were massed, designated M_1, M_2, M_3, M_4, M_5, and M_6; and two were distributed, designated D_1 and D_2. Each of these eight sets contained three words. Thus a total of $8 \times 3 = 24$ words were used in the experiment for a given *S*. Training sessions ran for 6 consecutive days. Each session used one of the M sets and one of the D sets. The M words were presented three times within a session, whereas the D words were presented once. Thus, there were $3 \times 3 = 9$ presentations of M items plus 3 presentations of D items yielding a total of 12 presentations in any one session. Words from a different M set were presented in each session and all the learning trials for the set occurred on the same day. Words from a given D set were presented on alternating days. Table 1 summarizes the daily presentations.

The arrangement of the list for the first training session (Day 1) illustrates the procedure used for the entire training sequence. The first four items of the day's list consisted of the three words in M_1 plus a randomly chosen word from D_1. The second four items consisted of the three M_1 words plus a second randomly chosen D_1 word. The last four items consisted of all three M_1 words plus the remaining word from D_1. In other words, the 12 presentations to an *S* on any day were given in three blocks with four words in a block. Each block contained all three M words and a randomly chosen D word. The order of the words within a block was randomly determined. Further, the assignment of words to M and D sets was completely counterbalanced over *Ss*, so that every word appeared equally often in the various M and D conditions.

Tests were administered 10 and 20 days after the end of the training sequence. The students did not receive any computerized drill between the training and test days. The basic test procedure consisted of presenting the complete list of 24 words. The order of the words for each *S* was randomly determined, and each word was presented once using the procedure of Figure 1. As during

TABLE 1

SUMMARY OF THE WORD SETS USED DURING THE SIX TRAINING SESSIONS

Condition	1	2	3	4	5	6
Massed (M)	M_1	M_2	M_3	M_4	M_5	M_6
Distributed (D)	D_1	D_2	D_1	D_2	D_1	D_2

the training sessions, the student was told whether or not his response was correct, and was then given 6 seconds to study the correct answer before the next item was presented.

RESULTS

Figure 2 presents the proportion of correct responses over successive presentations of M and D items. For example, on Day 1, the M_1 items were each presented three times; the proportions correct for each of the three presentations were averaged over *Ss* and plotted successively above Training Session 1. The D_1 items were each presented once; the mean proportion correct for these items is also plotted above Training Session 1. This was done for the data from each of the six training sessions. Approximately 2 minutes elapsed between two presentations of a massed item, whereas 2 days elapsed between any two presentations of a distributed item.

The tests were given on Days 16 and 26. The test results are also presented in Figure 2. The six massed curves are similar in form; they all rise sharply, then drop off by the time of the administration of the first test. In contrast, the two distributed curves rise more gradually but do not show a drop off at the time of the first test.

All items were presented three times during the training sequence and once on each of the test days. Figure 3 gives the proportion correct on each presentation averaged separately over M and D items. During the training sequence, the proportion correct for the M items increased from about .31 on the first presentation to .77 on the third presentation, whereas the D items correspondingly increased from about .25 to .57. The difference between the average proportion correct on the first presentation of M items and the first presentation of D

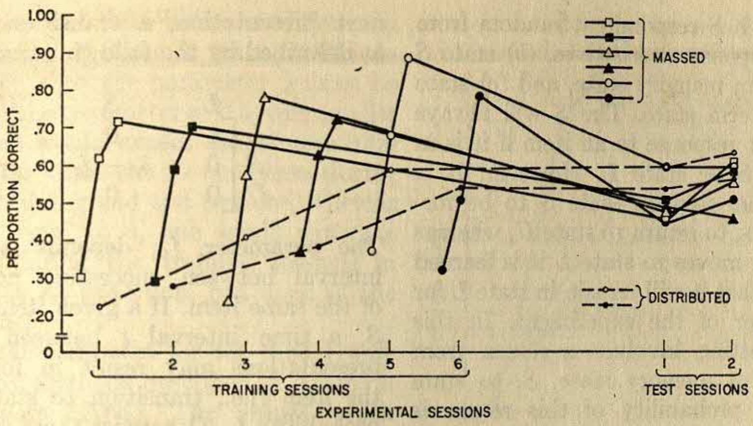


Fig. 2. Proportion of correct responses for massed and distributed items on both training and test trials.

items was not significant at the .05 level using a paired t test, $t = 1.58$, $df = 28$. However, there is no reason to expect equality when it is noted that the data point for the mean of the massed first presentations came from all six training sessions whereas the data point for the mean of the distributed first presentations came from the first two training sessions. In contrast, as indicated in Figure 3, there were significantly more correct responses on the second and third presentations of the M items than on the corresponding presentations of D items.

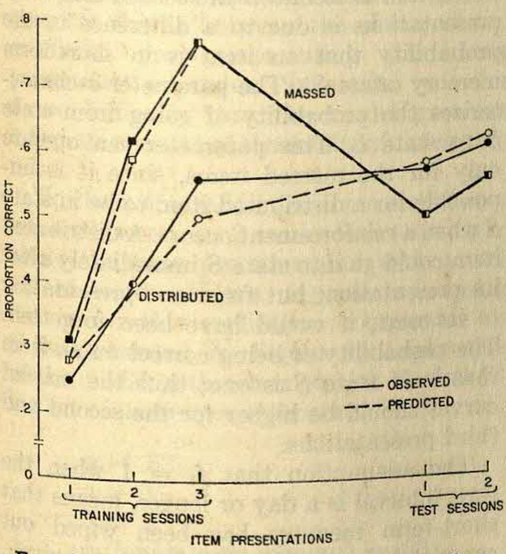


Fig. 3. Observed and predicted values for the massed and distributed conditions.

A paired t test on the combined data from the posttraining tests yielded $t = 2.44$, $df = 28$ which was significant at the .025 level, indicating that distributed practice resulted in better performance than massed practice.

Discussion

The major results of this experiment were: (*a*) the massed condition was superior to the distributed condition on the second and third presentations of the training sequence and (*b*) the distributed condition was superior on both of the test sessions. Thus, it appears that the massed repetitions are better if one looks at short-term performance, but in the long run more learning occurs when repetitions of an item are well distributed.

In this section, these data are analyzed in terms of a model that has been proposed to account for paired-associate learning. The model is a variation of the trial-dependent-forgetting model presented in recent articles by Atkinson and Crothers (1964) and Calfee and Atkinson (1965). The learning of a list of spelling words can be said to resemble the learning of a list of paired-associate items; no assumption is made that the two tasks are identical, yet there are variables in paired-associate learning that clearly are relevant to the spelling task.

In the model, S is assumed to be in one of three learning states with respect to a stimulus item: (*a*) state U is an unlearned

state, in which S responds at random from the set of response alternatives, (b) state S is a short-term memory state, and (c) state L is a long-term state. The S will always give a correct response to an item if it is in either state S or state L. However, it is possible for an item in state S to be forgotten, that is, to return to state U, whereas once an item moves to state L it is learned in the sense that it will remain in state L for the remainder of the experiment. In this model, forgetting involves a return from the short-term memory state, S, to state U, and the probability of this return is postulated to be a function of the time interval between successive presentations of an item.

More specifically, two types of events are assumed to produce transitions from one state to another: (a) the occurrence of a reinforcement, that is, the paired presentation of the stimulus item together with the correct response, and (b) the occurrence of a time interval between successive presentations of a particular item. The associative effect of a reinforcement is described by the following transition matrix:

$$\begin{array}{c} \\ L \\ S \\ U \end{array} \begin{array}{c} \begin{array}{ccc} L & S & U \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ a & 1-a & 0 \\ bx & (1-b)x & 1-x \end{bmatrix} \end{array}$$

Thus, if an item is in state U and the correct response is shown to S, then with probability $(1 - x)$ the item stays in state U, and with probability x the item moves into state S or L: If it moves, then with probability b it moves into L and with probability $(1 - b)$ into S. Similarly, if an item is in state S and the correct response is shown, then with probability a the item moves to state L, and with probability $1 - a$ the item stays in state S. Finally, if an item is in state L, then it remains there with probability 1. The parameter x is assumed to vary as a function of the familiarity of the items in the list being studied. Thus, during the test sessions involving 24 familiar items, x will be larger than during the initial study sessions involving 12 items, many of which are presented for the first time.

From one presentation of an item to its next presentation, a transition can occur as described by the following matrix:

$$\begin{array}{c} \\ L \\ S \\ U \end{array} \begin{array}{c} \begin{array}{ccc} L & S & U \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1-f_t & f_t \\ 0 & 0 & 1 \end{bmatrix} \end{array}$$

The parameter, f_t, depends on the time interval between successive presentations of the same item. If a given item is in state S, a time interval t between successive presentations may result in forgetting of the item (i.e., transition to state U) with probability f_t. Otherwise there is no change in state. For simplicity, we assume $f_t = 0$ for short time intervals within the range of a given training session. When the time interval is a day or greater, then we assume $f_t = 1$. In essence, no forgetting occurs from the short-term state within a given training session, but from one day to the next no information is retained in short-term store. Furthermore, the above transition matrices imply that L is an absorbing state; once an item enters state L it remains there. The model makes the additional assumption that at the start of the experiment an item is already known (state L) with probability p, or not known (state U) with probability $1 - p$.

For this model, the difference between the M and D items on the second and third presentations is due to a difference in the probability that an item is in short-term memory (state S). The parameter a characterizes the probability of going from state S to state L. This parameter can operate only for the massed items, since it is impossible for a distributed item to be in state S when a reinforcement occurs. A distributed item could go into state S immediately after its presentation, but from one presentation to its next, it would have been forgotten. The probability of being correct on an item that is in state S is one; thus the massed curves should be higher for the second and third presentations.

The assumption that $f_t = 1$ when the time interval is a day or longer, means that short-term memory has been wiped out completely by the time the first test is given. Thus, superiority of the D items over the M

items in the test data indicates differences in the number of items in state L. This in turn implies that the parameter b must be larger than the parameter a. If b were smaller than a, one would expect the M condition to do better than the D condition during both the training and test sessions, whereas if b were equal to a, one would expect a difference during the training sessions in favor of the M condition, but none in the test sessions.

Parameter estimates for the model were obtained by methods described in Atkinson and Crothers (1964). The values which yielded the best fit between observed and predicted proportions were:

$$p = .28$$
$$a = 0$$
$$b = .38$$
$$x \text{ (for training sessions)} = .45$$
$$x \text{ (for test sessions)} = .74$$

These estimates were consistent with the notion that b should be larger than a. The model proposed here is similar to Greeno's (1964) model for paired-associate learning in which he explicitly requires the parameter a to be zero. The present findings for this more complex task indicate that his theory and related research on paired-associate learning are relevant to the effect of repeated presentations of spelling items. Figure 3 presents the fit between the observed and predicted proportions using the above parameter estimates. Inspection of this figure indicates that the model gave an adequate account of the results of the experiment.

To check the validity of these results, the same S's were run 2 weeks later using precisely the same procedure but with a new set of words. Figure 4 presents learning curves for this replication comparable to those presented in Figure 3. Application of the model to this data yielded the following set of parameter estimates:

$$p = .32$$
$$a = 0$$
$$b = .33$$
$$x \text{ (for training sessions)} = .60$$
$$x \text{ (for test sessions)} = .72$$

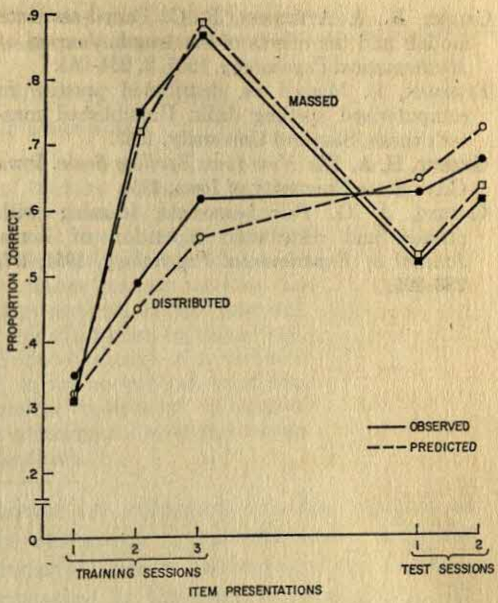


FIG. 4. Observed and predicted values for the replication experiment.

Once again, the estimate of a is zero confirming our earlier result. Also, in general, performance is superior in the second experiment, suggesting that some form of learning-to-learn may be operating in this situation.

The authors have not carried out analyses that bear on some of the more detailed features of the model. In fact, in view of the stimulus material used, it seems unlikely that these features would be verified. What clearly needs to be done is to generalize the paired-associate model to take account of the linguistic constraints imposed by the spelling task. Some of the present results and those of Knutson (1967) suggest guidelines for such a model but the authors are not prepared to be more specific at this time. Hopefully such a model would provide a more definitive answer to the problem of optimizing the instructional sequence in spelling drills.

REFERENCES

ATKINSON, R. C. Instruction in initial reading under computer control: The Stanford Project, *Journal of Educational Data Processing*, 1967, 4, 175–192.

ATKINSON, R. C., & CROTHERS, E. J. A comparison of paired-associate learning models having different acquisition and retention axioms. *Journal of Mathematical Psychology*, 1964, 1, 285–315.

CALFEE, R., & ATKINSON, R. C. Paired-associate models and the effects of list length. *Journal of Mathematical Psychology*, 1965, **2**, 254–265.

FISHMAN, E. Massed vs. distributed practice in computerized spelling drills. Unpublished master's thesis, Stanford University, 1967.

GREENE, H. A. *The New Iowa Spelling Scale*. Iowa City: State University of Iowa, 1954.

GREENO, J. G. Paired-associate learning with massed and distributed repetition of items. *Journal of Experimental Psychology*, 1964, **67**, 286–295.

KNUTSON, J. M. Spelling drills using a computer-assisted instructional system. Technical Report No. 112, Institute for Mathematical Studies in the Social Sciences, Stanford University, 1967.

SUPPES, P. The uses of computers in education. *Scientific American*, 1966, **215**, 207–220.

SUPPES, P., JERMAN, M., & GROEN, G. Arithmetic drills and review on a computer-based teletype. *Arithmetic Teacher*, 1966, April, 303–308.

(Received August 5, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 4, 297-301

EFFECTIVENESS OF FEEDBACK TO TEACHERS AS A FUNCTION OF SOURCE[1]

BRUCE W. TUCKMAN AND WILMOT F. OLIVER
Rutgers · The State University

286 teachers were separated by years of teaching experience and subjected to 1 of 4 conditions: (*a*) feedback from students only, (*b*) from supervisors, i.e., vice-principals only, (*c*) from both students and supervisors, and (*d*) from neither (no feedback). It was found that student feedback led to a positive change among teachers (as measured by change in students' ratings across a 12 wk. interval). Supervisor feedback added nothing to this effect when combined with student feedback, and when alone, produced change in a direction opposite to the feedback as compared to the no-feedback condition. Less experienced teachers showed greater receptivity to student feedback than their more experienced counterparts while the reverse held true for receptivity to supervisor feedback.

The problem of modifying the behavior of teachers is one that has been submitted to close scrutiny from a variety of vantage points. Techniques such as microteaching and the use of interaction process analysis have been employed, primarily with student teachers, as a means of altering their behavior. Underscoring the entire rationale for this approach, Daw and Gage (1967) recently said:

> It is highly plausible that feedback regarding how others feel about one's behavior will affect one's behavior. Whether this maxim will hold under a given set of practical circumstances must, however, be determined empirically [p. 181].

This study was an attempt to extend this "maxim" to conditions as yet untested.

Bryan (1963) has shown that teachers will alter their behavior as the outcome of receiving feedback from their students. The purpose of this study was to replicate Bryan's basic finding, using his instrument, and then to extend this finding by determining the relative effects of feedback from students and from supervisors (i.e., administrators responsible for instruction) on teachers' behavior. Moreover, Bryan's study did not include control over the variable of amount of teaching experience of teachers whose behavior was to be changed. His experimental and control groups showed an imbalance on this variable at the conclusion of his experiment with the preponderance of less experienced teachers appearing in the experimental group. An additional purpose of the present study was to systematically introduce years of teaching experience as an experimental variable so that its effects, if any, could be determined.

Finally, the present study was carried out with vocational teachers, in order to demonstrate additional generalizability for the basic finding obtained by Bryan using primarily teachers of academic subjects.

The fact that teachers change as the result of student feedback has also been demonstrated by Gage, Runkel, and Chatterjee (1960). Their study also showed that amount of change was related to the interval between pretest and posttest. Daw and Gage (1967) have shown, furthermore, that feedback from teachers can be used to alter the behavior of principals, but that the amount of change is not a function of the pretest-posttest interval.

In this study, as in previous studies in this area, the measurement of change in teacher behavior was inferential. Students were asked to rate their teacher twice, with a 12-week interval separating these ratings (during which time the treatments could take effect). Behavior change by teachers was inferred from a difference between postinterval and preinterval ratings. Remmers (1963) has shown that students, as a measuring instrument, are as

[1] This study was, in part, the doctoral dissertation of the junior author. It was supported, in part, by Grant No. 6-8327 from the United States Office of Education.

reliable as the best mental and educational paper-and-pencil tests and can discriminate between aspects of teacher behavior (see also Tuckman, 1967). Thus, the dependent variable was identified as change in teachers' behavior with the recognition that this was inferential.

The expectation that years of teaching experience would be a significant variable was based on studies such as that of Ryans (1964) and Peterson (1964) who have shown that teachers' behavioral patterns change in a systematic fashion as a function of age. While age and years of teaching experience are not the same variable, they are assuredly related, with the latter being perhaps the more conceptually meaningful in an educational context.

PROBLEM

To determine the relative effects of students and supervisors as feedback sources for teachers, four conditions were run. In the first condition student feedback alone was employed; in the second, supervisor feedback was employed alone (the supervisor being an administrator, usually a principal or vice-principal responsible for the teaching activities of teachers); in the third, both feedback sources were employed concomitantly; and in the fourth, no feedback was given. Teachers were further classified as to teaching experience and systematically assigned to conditions on that basis.

It was hypothesized that: (*a*) teachers receiving feedback would change more than teachers not receiving feedback (essentially a replication of Bryan's results); (*b*) amount of change in teachers' behavior would vary as a function of feedback source; (*c*) years of teaching experience and amount of change would be inversely related.

METHOD

Sample

The sample consisted of 286 teachers of vocational subjects at the high school or technical institute level. Schools were selected from New Jersey and surrounding out-of-state counties and virtually all the vocational teachers in the schools used took part in the study. Participating teachers had a median class size of 15 students who were either in the tenth, eleventh, or thirteenth grade.

Measurement of Teacher Behavior

Teacher behavior was measured by the Student-Opinion Questionnaire (SOQ) developed by Bryan (1963). This instrument includes 10 rating scales on which teacher is judged as to his (*a*) knowledge of his subject, (*b*) ability to explain, (*c*) fairness, (*d*) ability to maintain discipline, (*e*) degree of sympathetic understanding, (*f*) ability to make you learn, (*g*) ability to be interesting, (*h*) ability to get things done efficiently, (*i*) ability to get students to think for themselves, and (*j*) general all-round teaching ability. Each scale has five points labeled: below average, average, good, very good, and the very best.

Bryan (1963) has reported reliability coefficients for the 10 items on the SOQ of from .75 to .85 for chance-half averages for 50 classes. For whole classes of 28 students on the average, coefficients of from .86 to .92 were obtained.

On the reverse side of the SOQ are four open-ended questions dealing with the course and teacher, reflecting on things that are liked about each and suggestions for the improvement of each.

Feedback Conditions

Students only. Students completed the SOQ, and their ratings on the 10 scales were averaged. The teacher was presented with a graph showing the average student judgment for each item. In addition, a summation of the students' responses on the open-ended questions were provided. Teachers were told that the feedback was from their students.

Supervisor only. The teacher's supervisor (either the principal, vice-principal, or assistant principal) completed the SOQ, and his ratings on each item were given to the teacher in graphical form along with a summary of his answers to the open-ended questions. The teacher was told that this rating was made by the supervisor. (In this condition, student ratings were also obtained although these were not made available to the teacher.)

Students and supervisor. The teacher's supervisor and students completed the SOQ, and feedback from each was given separately, along with identification of source in the same manner as in the first two conditions.

No feedback. Students completed the SOQ, but no feedback was provided to the teacher.

All initial testing was done in the late fall.

Years of Teaching Experience

Based on information from a personal information form, teachers were categorized as having 1–3 years of teaching experience, 4–10 years of teaching experience, or 11 or more years of teaching experience. Teachers from each group were then randomly assigned to each condition. The

TABLE 1

DESIGN OF THE EXPERIMENT: ASSIGNMENT OF TEACHERS TO TREATMENT AND EXPERIENCE GROUPS

Condition	Years of experience of instructor					
	1–3 years (A_1)		4–10 years (A_2)		11 or more years (A_3)	
	B_1	B_2	B_1	B_2	B_1	B_2
No student feedback (C_1)	14	18	19	18	18	13
Student feedback (C_2)	39	32	25	31	32	27

Note.—Cell entries are number of observations per cell; $N = 286$; Abbreviated: B_1 = no supervisory feedback, B_2 = supervisory feedback.

overall design of the study and assignment of teachers to conditions is shown in Table 1.

Measurement of Change in Teacher's Behavior

In the late spring, following a 12-week interval after the initial testing, students of each of the teachers in the study completed the SOQ. The measure of change in each condition was the sum of the differences between the preinterval judgments by the students on the 10 items and their postinterval judgments. Ratings on each item were averaged across students and the preinterval average on each item was then subtracted from the postinterval average to yield a change score on each of the 10 items. These 10-item change scores were summed to obtain a total change score. Student judgments were used throughout as a measure of change to maintain a constant measuring instrument across conditions. This was seen as justifiable since preinterval ratings by students did not differ significantly from those of supervisors in conditions where both were obtained and the latter were used as the feedback source.

All test administration was accomplished by the local vocational guidance counselor.

Analysis

For purposes of analysis, the four feedback conditions (Conditions 1–4) were treated as two factors: supervisor feedback and student feedback, with two levels on each: present and absent. The four conditions were thus labeled as follows: (b_1c_1) student and supervisor feedback, (b_1c_2) supervisor feedback only, (b_2c_1) student feedback only, and (b_2c_2) no feedback (see Table 1). Years of teaching experience was the first factor and had three levels. Subsequently, a 3 × 2 × 2 analysis of variance using the unweighted means solution for unequal cell entries (Winer, 1962) was carried out on the total change score for each teacher. (Each teacher was used only once in the design.) In addition, direct mean comparisons were made using the Duncan multiple-range test (Duncan, 1955).[2]

RESULTS

The results of the analysis of variance for the total change score showed that the presence of student feedback (Factor C) had a significant effect on teachers' behavior as compared to its absence ($F = 5.941$; $df = 1/274$, $p < .025$) while the presence of supervisor feedback (Factor B) produced no significant effect ($F = 1.064$; $df = 1/274$). The years-of-experience variable (Factor A) also failed to produce a significant effect ($F = 0.701$; $df = 2/274$) and none of the interactions achieved significance at the .05 level ($F < 1$ in each case).

In an effort to delineate further the feedback effects, means for the four feedback conditions were compared, as shown in Table 2. From the table it can be seen that both conditions involving student feedback showed significantly greater change than both conditions not involving student feedback.[3] Feedback from students

TABLE 2

MEAN TOTAL CHANGE SCORES BY FEEDBACK CONDITION AND THEIR COMPARISON BY DUNCAN MULTIPLE-RANGE TEST

Students only	Students and supervisors	Supervisors only	No feedback
−.054	−.385	−2.449*	−1.234*

* Significantly different from all other means, $p < .01$ (with exception of difference between second and fourth means, where $p < .05$).

[2] A fifth condition, called the posttest-only control group by Daw and Gage (1967), was also run with an additional 15 teachers. These teachers were rated by their students only at the end of the interval. The purpose of this condition was to determine whether the pretest or preinterval measurement had a sensitizing effect on the raters or teachers (cf. Campbell & Stanley, 1963). A comparison of the mean for this posttest-only control group to the mean on the postinterval measurement for the no-feedback group showed them to be comparable. Thus, it was concluded that test sensitization was not a source of invalidity.

[3] Throughout this description, results are referred to as changing "more" or "less." However,

alone and from students and supervisors combined were statistically comparable, indicating a failure for feedback from supervisors to generate any change beyond that accounted for by student feedback alone. Finally, feedback from supervisors alone produced a significantly greater negative shift (i.e., a change in the opposite direction of that recommended by the feedback) than no feedback at all.

Thus, student feedback "improved" teacher behavior as compared to no feedback. Supervisor feedback produced no additional effect when combined with student feedback, and an adverse effect when used alone.

DISCUSSION

The first hypothesis of this study predicted that feedback (source unspecified) would yield a greater positive change than no feedback, while the second hypothesis predicted different effects for the different feedback sources. The surprising finding of this study was that teachers receiving feedback from supervisors changed more in the opposite direction from the feedback than the spontaneous shift obtained in the no-feedback condition. Thus, the first hypothesis holds true for student feedback (a replication of Bryan's findings) which led to effects in excess of the no-feedback condition. Supervisory feedback added nothing to the student feedback effect when they were combined. (If anything it reduced it, but not significantly so.) Since supervisory feedback had the opposite effect than predicted, the second hypothesis was confirmed—that is, the feedback sources did have different effects. If in the first hypothesis, it was simply predicted that feedback would produce greater changes than no feedback, it would have been confirmed. Certainly this experiment suggests that teachers react to feedback, irrespective of source, with these reactions being positive only in the case of student feedback.

The question of why teachers reacted to feedback from supervisors as they did is immediately raised. It can only be surmised that teachers are defensive toward (or even hostile to) administrators who, in the absence of much basis for judgment, attempt to tell them how to teach. Of interest, though, is the fact that within the educational milieu, the only source of feedback to teachers, typically, are their supervisors. The data collected here indicate that such feedback is doing more harm than good, with the "best" source of feedback, students, overlooked.

The third hypothesis of the present study predicted an inverse relation between years of experience and receptivity to feedback. While the obtained relationship was not sufficiently strong to prove significant, the most experienced teacher group tended to show the least receptivity to feedback from their students, as the hypothesis predicted. However, the least experienced teacher group tended to show the least receptivity (i.e., the least relatively positive shift) to feedback from their supervisor—the reverse of the hypothesis.

Finally, a last question must be raised. Why do all the change scores tend to be negative with positive change being measured in terms of the "smallness" of the negative score? The use of a group of teachers whose students made only the postinterval ratings indicated that the test-retest phenomenon was not responsible for this shift from pre- to postratings. It appeared that students are more critical of their teachers at the end of the term than at the middle. At the time when the teacher is about to evaluate and grade the student, the student perhaps replies in

---

in the light of the fact that almost all of the means are negative, changing more means showing a lesser negative shift (i.e., a smaller negative change score) while changing less means showing a greater negative shift (i.e., a larger negative change score). This tendency for ratings to be less positive following the interval as compared to those preceding the interval were not attributable to a testing effect (see the preceding footnote). One must conclude that students as raters are more negatively inclined toward their teachers in the spring (after experiencing them for a year) than in the fall. Thus, the positive effect of feedback, when it occurred, was to reduce this tendency toward greater negativity of ratings (i.e., make the negative score smaller or positive).

kind. Thus, a positive change appeared as a lessening in the "naturally" occurring negative shift. Researchers interested in using student judgments are cautioned to use the same starting and ending times for all groups to avoid the confusion of this end-of-term effect. September to January will not lead to the same effect as February to June.

REFERENCES

BRYAN, R. C. Reactions to teachers by students, parents, and administrators. United States Office of Education, Cooperative Research Project No. 668. Kalamazoo: Western Michigan University, 1963.

CAMPBELL, D. T., & STANLEY, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally, 1963.

DAW, R. W., & GAGE, N. L. Effect of feedback from teachers to principals. *Journal of Educational Psychology*, 1967, **58**, 181–188.

DUNCAN, D. B. Multiple range and multiple F tests. *Biometrics*, 1955, **11**, 1–42.

GAGE, N. L., RUNKEL, P. J., & CHATTERJEE, B. B. Equilibrium theory and behavior change: An experiment in feedback from pupils to teachers. Report No. 6 in Studies in the generality and behavioral correlates of social perception. Urbana: Bureau of Educational Research, College of Education, University of Illinois, 1960.

PETERSON, W. A. Age, teacher role, and the institutional setting. In B. J. Biddle & W. J. Ellena (Eds.), *Contemporary research on teacher effectiveness*. New York: Holt, Rinehart & Winston, 1964.

REMMERS, H. H. Rating methods in research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally, 1963.

RYANS, D. G. Characteristics of teachers. In B. J. Biddle & W. J. Ellena (Eds.), *Contemporary research on teacher effectiveness*. New York: Holt, Rinehart & Winston, 1964.

TUCKMAN, B. W. *A study of the effectiveness of directive versus non-directive vocational teachers as a function of student characteristics and course format*. United States Office of Education, Project No. 6-2300, Progress Report No. 1. New Brunswick: Rutgers. The State University, 1967.

WINER, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

(Received August 14, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 4, 302-308

EFFECT OF SOCIAL PRESSURE ON CONCEPT IDENTIFICATION[1]

VERNON L. ALLEN AND BARRY W. BRAGG
University of Wisconsin

This study investigated the effect of veridical and nonveridical group feedback on concept identification, and the transfer of the effect of such social pressure from 1 problem to a 2nd one. Social pressure consisted of a group of Ss giving either unanimously correct or incorrect responses over a series of trials. To study transfer of the social pressure effect, in 1 condition the group gave veridical feedback on the 1st problem and nonveridical feedback on the 2nd; the opposite order of feedback was given in another condition. Results showed that veridical group feedback facilitated concept acquisition and that nonveridical feedback depressed acquisition. Moreover, transfer of the social pressure effect occurred between the 2 problems, resulting in poorer performance on the 2nd problem.

Research during the past decade has shown conclusively that social pressure from a group influences individual behavior on a variety of simple judgmental tasks, for example, perceptual discriminations (Allen, 1966). Yet, little research has been directed toward investigating the possible role of social pressure in more complex cognitive processes such as learning and remembering. The paucity of research on the effect of social pressure on complex behavior is perhaps due to the tendency of social psychologists to consider the nature of the task unimportant or irrelevant in comparison to the basic psychological processes under investigation. Thus, a task is often employed solely because it is simple and available; many such tasks no doubt tap only simple psychological processes. Because of this emphasis on processes rather than tasks, the study of social pressure has largely neglected the investigation of complex cognitive behavior.

Our knowledge concerning the effects of social pressure on behavior indicates that the complex cognitive processes of learning and remembering might be particularily vulnerable to social influence at certain stages of learning. For example, the literature on social pressure shows that effects of the group are more pronounced when the task is ambiguous (Luchins, 1945; Walker & Heyns, 1962) or when the person has little confidence in his ability to make a correct response (Hochbaum, 1954; Wiener, 1958). During the initial phases of the learning process the task is quite ambiguous to *S*, and his confidence in his ability to respond correctly is low. At this stage, it is very likely that social pressure would exert a strong influence on learning; the effect of such social pressure could, of course, aid or hinder the speed of learning, depending on the objective correctness or incorrectness of the group's response.

Little research has been conducted on the effect of social pressure on learning and remembering. Allen and Bragg (1967) showed that social pressure influences memory on a paired-associated learning task. One study of acquisition (Rhine, 1960) employed a very simple learning situation in which *S*s were asked to predict whether a "little known" group of people possessed each of a series of desirable and undesirable traits. Results showed that peer-group responses aided acquisition on this simple task. In view of the meager systematic data available, the first purpose of the present study was to explore the effect of social pressure during the acquisition phase of learning. Social pressure is presented in the present study in the form of

[1] The research reported herein was performed pursuant to a contract with the United States Office of Education, Department of Health, Education, and Welfare under the provisions of the Cooperative Research Program.

unanimous (correct or incorrect) feedback from a group of S's peers. In order to avoid the limitations of the simple rote learning situation, the concept identification task was chosen for use in this study.

A second purpose of the present study was to investigate transfer of the effect of social pressure from one problem to another. Insufficient attention has been devoted to the potential sequential effects of group pressure. A few studies have addressed themselves to the problem in a very limited way by examining between-trial effects of social pressure on a single task. One study found a carry-over of social pressure between trials on perceptual judgments of numerosity of a pattern of dots (Fisher, Rubenstein, & Freeman, 1956). In this study a confederate consistently gave estimates higher than S's estimates. Not only was S influenced by the confederate's response to the same stimulus, but S's initial response on the next stimulus display—given prior to the confederate's estimate—was also affected. That is, between-trial influence as well as within-trial influence was demonstrated. In a subsequent study, Peterson, Saltzstein, and Ebbe (1967), again using numerical estimates of dots, found between-trial influence when the stooge changed his response each trial in order to maintain a constant discrepancy from S's preceding estimate. But when the stooge maintained a fixed absolute estimate, no between-trial influence was observed.

Relevant to the question of sequential effects of group pressure is Hollander's (1960) research, indicating that tolerance of an individual's deviation from the group is a function of his earlier behavior in relation to the group. Greater acceptance of an individual's attempt to change the group norm was shown when the individual's conformity to the group occurred in the earlier stages of interaction, rather than at later stages.

When the direction of the group's response changes over time, sequential effects become crucial; the possibility of transfer effects from one task to another then arises. A task having an objectively correct answer, such as the concept attainment task, would appear to possess distinct advantages for studying transfer effects of group pressure. Use of such a task allows us to shed some light on the question of appropriateness or efficiency of conformity and nonconformity. Much controversy exists concerning whether conformity to the group should be considered desirable or undesirable behavior. Under certain circumstances, conformity to a group is undoubtedly a very adaptive and appropriate response. To agree with a group that gives veridical or objectively correct responses in a concept identification task, and to depend upon the group when one is uncertain, would facilitate learning. By contrast, if the group's responses were nonveridical or incorrect, to agree because of social pressure is clearly inefficient since it would interfere with learning. The study of transfer effects of group pressure has been neglected, but the phenomenon attains considerable importance when dealing with objective tasks.

Social reality is complex; the behavior of a group does not always remain consistent over time. Agreement with a group is therefore advantageous to the individual in some circumstances and disadvantageous in others. Consider the responses of a group on a concept identification task. As pointed out earlier, agreement with the group would facilitate identification of the concept if the group supplied correct responses. Suppose that the group's responses were initially correct, and that S came to rely on the group. If the same group later began giving incorrect responses S's continued reliance on the group would hinder learning by delaying prompt adaptation to the new situation. Negative transfer effects of two types are therefore possible: (*a*) initial conformity to a correct group, followed by later conformity to the same group now giving incorrect responses, (*b*) initial nonconformity to an incorrect group, followed by nonconformity to the group now giving correct responses.

Ideally, an individual's behavior would consist of a high degree of selectivity in relation to the group. Because the group's response is subject to change, selective dependence, rather than rigid conformity or

nonconformity, is most advantageous to the individual. Therefore, the most efficient relation of the individual to the group is conformity when the group is correct and nonconformity when the group is incorrect.

In summary, the purposes of this study are twofold: First, to investigate the effect of social pressure on concept identification; and second, to study the transfer of the social pressure effect from one task to another.

Method

Subjects

The *S*s for the study were 73 female freshman and sophomore students who volunteered to participate in the experiment without compensation of any kind. The *S*s were randomly assigned to the five experimental conditions.

Apparatus

The apparatus, a Crutchfield (1955) electrical signaling device used in conformity research, consists of five booths containing nine response switches and a matrix of 45 signal lights showing the answers given by the five *S*s. Modification of the apparatus by relabeling switches permitted its use in the concept attainment task. The *S* is led to believe that she responds last in a group of five, and that other persons' responses are shown in each booth. In actuality, all five *S*s answer in the last position, and the signal lights shown in each *S*'s booth are controlled by *E* from another room. In this way, all *S*s are exposed to the same pattern of group pressure, and stooges are unnecessary. Five *S*s were always tested together.

Material

The learning task consisted of slides containing various geometric designs. The slides were projected on a screen 10 feet in front of *S*s. Five dimensions, varying on two attributes, were used in the concept attainment task. The dimensions were: (*a*) size—large or small, (*b*) shape—square or circle, (*c*) color—red or green, (*d*) number—one or two, and (*e*) texture—plain or textured.

Instructions

The *S*s were told that their task was to solve a concept identification problem. A slide containing all five dimensions was first shown and the dimensions were described. The *S*s were told that the concept to be identified would consist of one or a combination of the dimensions present in each slide. The *S*s' task was to determine whether or not a slide contained the concept, and to identify the relevant dimensions.

The *S* was further told in the instructions:

> If you think that the slide does contain the concept, you should turn on switch number one marked "contains concept." If you think that the slide does not contain the concept, turn on switch number two marked "does not contain concept." If you have decided the slide contains the concept, then I want you to turn on one or more of the five switches which indicates the relevant or correct dimension. For example, if on this slide you believed that the correct concept was "small green circle," you would first turn on switch number one, then the switches corresponding to size, color, and shape (switches 5, 6, and 7). If you thought that this slide did not contain the concept, and therefore turned on switch number two, I want you to turn on the switch or switches for the dimension or dimensions that are incorrect. For example, if you thought that the concept was "small red circle" (instead of "small green circle"), then the incorrect dimension would be color and you would turn on the switch corresponding to color (switch number 5). After you all have answered, I will tell you if the slide contained or did not contain the concept. Then we will go on to the next slide.

Five practice trials were given, followed by 25 test trials in each of the two series of trials. In the group conditions, *S*s were assigned a position for responding and always answered in order. When the experiment began all *S*s, unknowingly, were assigned to the last position, Number 5. This allowed *E* to control the simulated responses observed by *S* prior to her answering. The first practice slide was always an example of the concept ("red-textured" for the first series, and "small" for the second series).

In summary, *S*s responded by pressing one switch to indicate that a slide contained the concept, or a second switch to indicate that the slide did not contain the concept. One or more of five other switches in *S*'s booth, labeled by dimension (color, shape, etc.), were used by *S* to indicate correct dimensions if the slide contained the concept or to indicate incorrect dimensions if the slide did not contain the concept. After each trial, *E* reported whether or not the slide contained the concept, but he did not give information concerning the correctness of the dimensions comprising the concept.

Design

The five conditions used in the study are described below. In each condition *S* received two concept identification problems, each problem consisting of a series of 25 trials.

Control. In the control condition, *S*s learned the two concept attainment tasks without seeing the other five *S*s' answers to the 25 slides in each series. Twenty-one *S*s were used in this condition.

Veridical. In this condition, the feedback *S*s received was mostly correct for each of the two concept attainment tasks. On the first 12 slides of the first series of 25 trials, there was some dis-

agreement shown among the simulated *S*s in order to increase credibility of the situation. But after the 12th slide the simulated *S*s appeared unanimously to choose the correct concept, and adhered to the concept for the remainder of the series. The same sequence of trials was used for supplying feedback in Conditions 4 and 5 below. Fifteen *S*s were used in this condition.

Nonveridical. The responses of others that *S* observed in this condition were incorrect on both problems. On the first 12 slides there was disagreement shown by the simulated *S*s in their incorrect responses. But subsequent to Slide 12 all *S*s agreed on the concept, giving identical wrong answers for each slide. The same sequence of incorrect trials was used for incorrect feedback in Conditions 4 and 5 below. Thirteen *S*s were used in this condition.

Veridical-nonveridical. In this condition *S*s received correct feedback from the group on the series of 25 trials comprising the first problem, and incorrect group feedback on the second 25 trials for the second problem. Eleven *S*s were used in this condition.

Nonveridical-veridical. In this condition *S*s received incorrect group feedback on the first problem, but correct group feedback on the second problem. Thirteen *S*s were used in this condition.

RESULTS

The most straightforward overall analysis of the data consists of calculating the percentage of *S*s in each feedback condition who correctly identified the concept used in the two problems by the end of each series of 25 trials. For this analysis data will be combined for the two veridical feedback conditions, and for the two nonveridical conditions.

Table 1 presents results for the first concept identification problem. Data in Table 1 show that on the first task 88% of the *S*s in the veridical feedback condition correctly identified the concept, as compared with 23% in the nonveridical feedback condition. The difference between the veridical and nonveridical feedback conditions was statistically significant at less than the .01 level ($\chi^2 = 21.98$, $df = 1$). In the control condition, where *S* could not observe other persons' responses, 43% of the *S*s identified the concept. In the veridical feedback condition *S*s' performance was significantly better than in the control condition ($\chi^2 = 11.11$, $df = 1$, $p < .01$); but the decrease in performance of *S*s in the nonveridical feedback condition, relative to the control, was not statistically significant ($\chi^2 = 2.09$, $p < .20$).

TABLE 1

CONCEPT ATTAINMENT ON THE FIRST PROBLEM

Condition	*N*	Correct	Incorrect
Control	21	43	57
Veridical feedback	26	88	12
Nonveridical feedback	26	23	77

Note.—Entries are percentage of *S*s in each condition who correctly identified the concept, and the percentage who failed to do so.

Results for the second concept identification problem were congruent with results for the first problem. It can be seen in Table 2 that veridical feedback from the group improved performance while nonveridical feedback depressed performance. In the veridical feedback condition, 79% of the *S*s correctly identified the concept, as compared with only 12% in the nonveridical condition. Results for the control condition fell between the two experimental conditions (43%). The difference between the veridical and nonveridical feedback conditions was statistically significant at less than the .01 level ($\chi^2 = 22.59$, $df = 1$). In addition, scores in the veridical feedback condition were significantly better than in the control condition ($\chi^2 = 6.58$, $df = 1$, $p < .02$), and scores in the nonveridical feedback condition were significantly poorer than in the control condition ($\chi^2 = 4.56$, $df = 1, p < .05$).

In summary, the highly significant differences observed as a function of type of group feedback indicate that social pressure affected concept attainment, with veridical group responses facilitating per-

TABLE 2

CONCEPT ATTAINMENT ON THE SECOND PROBLEM

Condition	*N*	Correct	Incorrect
Control	21	43	57
Veridical feedback	28	79	21
Nonveridical feedback	24	12	88

Note.—Entries are percentage of *S*s in each condition who correctly identified the concept, and the percentage who failed to do so.

formance and nonveridical responses interfering with performance.

A second problem of interest in this study was the transfer of the group's effect on concept attainment from the first task to the second one. Recall that in the veridical-nonveridical condition, the group gave correct responses on the first task, but incorrect answers on the second task. The opposite inconsistent order of group feedback was followed in the nonveridical-veridical condition. The remaining two experimental conditions, in which the direction of the group's responses remained consistent across the two problems, provided a base line against which transfer of the group's inconsistent feedback across problems could be assessed.

Results showed that transfer effects were clearly evident. Data were first examined for the two groups that received veridical feedback on the second task. Feedback in one of these conditions (veridical) was also correct on the first task, but in the other condition (nonveridical-veridical) feedback was incorrect on the first task. One would predict that transfer of the effects of incorrect group feedback given on the first task would detrimentally affect concept learning on the second task; in other words, negative transfer should occur. Results showed that the mean trial on which the concept was correctly identified when both tasks received veridical group feedback was 15.2, as compared with 17.7 when the first task had received nonveridical group feedback. The difference between the two conditions was significant at beyond the .05 level of confidence by the one-tailed t test ($t = 1.98$).

To measure the transfer effect of veridical group feedback when the second task received incorrect feedback, it was necessary to analyze the data somewhat differently. Using the mean trial on which the concept was attained was not feasible because so few *S*s correctly identified the concept when incorrect feedback was given on the second task. Therefore, the mean number of times *S*s agreed with the incorrect responses of the group on the second task was used as an index of the transfer effect. Performance on the second task receiving nonveridical group feedback was then analyzed as a function of whether the group's feedback had been veridical or nonveridical on the first task. It was predicted that agreement with the group's incorrect responses on the second task would be higher when feedback from the group on the first task had been correct than when group feedback on the first task had been incorrect. Results supported the prediction: The mean number of trials on which *S*s agreed with the incorrect responses of the group on the second task was 11.5 for the veridical-nonveridical condition, as compared with a mean of 7.5 for the nonveridical condition. The difference between the two conditions was significant at the .05 level by a one-tailed t test ($t = 1.82$).

DISCUSSION

Results of the present study have shown that social pressure, in the form of unanimous responses of a group of peers, significantly affects behavior on a concept identification problem. The strength of the effect of social pressure on concept acquisition appears to be asymmetrical. Interestingly and, perhaps encouragingly, the amount of the facilitating effect of social pressure in the form of veridical group feedback was approximately twice as great as the amount of the detrimental effect due to nonveridical group feedback. The greater effect of correct feedback than of incorrect feedback is in accord with a study by Jones, Wells, and Torrey (1958), in which the *E* provided objective feedback to the group.

It should be emphasized that the amount of facilitation of concept acquisition attributable to veridical group feedback was not insubstantial. In the veridical feedback condition 89% of the *S*s accurately identified the concept as compared with 43% in the control condition, an advantage of 46% attributable to the group's feedback.

Whether the effect of social pressure was due to mere public agreement with the group or represented the individual's true belief is difficult to determine with certainty. The problem-solving situation is one that would primarily tap informational rather than normative social in-

fluence (Deutsch & Gerard, 1955). That is, agreement with the group was probably due to *S*'s using the responses of other persons as reliable sources of information about a solution to the problem, rather than to an attempt on *S*'s part to gain approval or avoid disapproval from the other group members. Instructions concerning the experiment and the nature of the task both served to orient *S*s toward utilizing other members of the group as informational rather than normative sources of influence. So it is very plausible to interpret the influence of the group as being due primarily to informational influence, and not to mere public compliance to the group.

It is interesting that although the task was equally unfamiliar to all group members, *S* was willing often to agree with the answers given by the group. No doubt such agreement served to reduce *S*'s motivation to search for the solution to the concept identification problem. Unanimity among a group of persons often means that their responses are correct; initial acceptance of such an assumption probably led *S* to place undue dependence on the group. Agreement with an apparently self-confident group could have caused a decrease in cognitive arousal on the part of *S*. As a consequence, *S* probably relaxed somewhat and exerted less cognitive effort in finding the solution to the problem. Such relaxation is perhaps also partially due to *S*'s acquiring a "set" of agreeing with the group which is difficult to break. Like *S*s in Luchin's (1942) water-jar problem, once the set is established, a new and critical analysis of the problem is accomplished slowly and with difficulty. The cognitive set to agree with the group can sometimes clearly aid *S*'s problem-solving attempts or equally often serve as a barrier, depending on the degree of veridicality of the group's responses.

Evidence of transfer of the group's effect on concept acquisition from the first to the second task is a very intriguing finding. When the group had previously given correct responses on the first task, *S*s were likely to continue agreeing with the group on the second task, although the group now gave incorrect responses. Conformity to the group at this time was inappropriate because the group's behavior was inconsistent with its previous veridical responses. Similarly, *S*s were unnecessarily inefficient when the group had given incorrect responses on the first problem, but changed to supplying correct answers on the second; in this case, *S*s conformed less than was warranted by the group's veridical answers given at that time.

The answer to the value question of whether conformity is desirable or undesirable obviously is shown in this study to depend on the specific characteristics of the situation. Appropriate and efficient behavior would consist of an individual's conforming to a group sometimes on some issues, and disagreeing at other times on other issues. The difficulty of increasing selective response to group pressure is, however, very real. The tendency is strong to respond consistently to a group (or person or situation), even though rational analysis would dictate a change in response. The transfer phenomenon observed in the present experiment appears to be a special case of a more general psychological phenomenon found in other contexts. For example, the halo effect observed in prestige suggestion is a case of behaving consistently toward an individual across situations, though a change in behavior would be the more reasonable response (Aronson & Golden, 1962). A source having high prestige on one topic tends to produce unwarranted agreement on another topic on which he has little competence; similarly, a source having low prestige on one topic often produces lower agreement than warranted on a second topic on which he has limited competence.

Transfer effects of the type found in the present experiment are probably not uncommon in everyday social behavior, but the problem remains to be systematically explored in future research.

REFERENCES

ALLEN, V. L. Situational factors in conformity. In L. Berkowitz (Ed.), *Advances in experimental social psychology*. New York: Academic Press, 1966.

ALLEN, V. L., & BRAGG, B. W. Effect of group pressure on memory. *The Journal of Psychology*, 1968, **69**, 19–32.

ARONSON, E., & GOLDEN, B. D. The effect of relevant and irrelevant aspects of communicator credibility on opinion change. *Journal of Personality*, 1962, **30**, 135–146.

CRUTCHFIELD, R. S. Conformity and character. *American Psychologist*, 1955, **10**, 191–198.

DEUTSCH, M., & GERARD, H. B. A study of normative and informational influences upon individual judgment. *Journal of Abnormal and Social Psychology*, 1955, **51**, 629–636.

FISHER, S., RUBENSTEIN, I., & FREEMAN, R. W. Intertrial effects of immediate self-committal in a continuous social influence situation. *Journal of Abnormal and Social Psychology*, 1956, **52**, 200–207.

HOCHBAUM, G. M. The relation between group members' self-confidence and their reactions to group pressure to conformity. *American Sociological Review*, 1954, **19**, 678–687.

HOLLANDER, E. P. Competence and conformity in the acceptance of influence. *Journal of Abnormal and Social Psychology*, 1960, **61**, 361–365.

JONES, E. E., WELLS, H. H., & TORREY, R. Some effects of feedback from the experimenter on conformity behavior. *Journal of Abnormal and Social Psychology*, 1958, **58**, 207–213.

LUCHINS, A. S. Mechanization in problem solving. *Psychological Monograph*, 1942, **54**(6, Whole No. 248).

LUCHINS, A. S. Social influences on perception of complex drawings. *Journal of Social Psychology*, 1945, **21**, 257–273.

PETERSON, D. E., SALTZSTEIN, H. D., & EBBE, C. Sequential effects in social influence. *Journal of Personality and Social Psychology*, 1967, **6**, 169–174.

RHINE, J. The effect of peer group influence upon concept-attitude development and change. *Journal of Social Psychology*, 1960, **51**, 173–179.

WALKER, E. L., & HEYNS, R. W. *An anatomy for conformity*. Englewood Cliffs, N. J.: Prentice-Hall, 1962.

WIENER, M. Certainty of judgment as a variable in conformity behavior. *Journal of Social Psychology*, 1958, **48**, 257–263.

(Received August 18, 1967)

JOURNAL OF
EDUCATIONAL PSYCHOLOGY

Volume 59 | October 1968 | Number 5

CONTENTS

© 1968 by the American Psychological Association, Inc.

Journal of Educational Psychology
1968, Vol. 59, No. 5, 309-314

SUCCESSIVE VERSUS CONCURRENT PRESENTATION OF MULTIPLE GRAPHEME-PHONEME CORRESPONDENCES[1]

JOANNA P. WILLIAMS
University of Pennsylavania

LIBRARY

Many new instructional programs recommend that the beginning reader be given material built on a simplified, regularized pattern of 1:1 grapheme-phoneme correspondences. To test this assumption, 2 methods of training multiple correspondences (1 grapheme mapping to 2 phonemes) were compared in a paired-associate paradigm. In successive training, only 1 of the 2 phonemes associated with a particular grapheme was presented at a time, while in concurrent training, both phonemes associated with each grapheme were introduced and practiced concurrently. Results suggest that concurrent training is superior, both in terms of the kind of "set" developed and of the performance level on the correspondences given in training. These findings run counter to the typical recommendations.

One of the fundamental concepts that must be developed by the beginning reader of English—or of any language based on the alphabetic principle—is that there is a correspondence between spoken language and orthography. There is some question as to what the critical unit of correspondence actually is, for each individual letter does not necessarily correspond to a phoneme (Gibson, Gibson, Pick, & Ossen, 1962; Hall, 1961). While recent analyses suggest that English orthography is a more regular—and more complex—system than had heretofore been recognized (Weir & Venezky, 1965), nevertheless, even when combinations of graphemes (spelling patterns) are considered as the basic unit, there is not a one-to-one correspondence. The irregularities add considerably to the difficulty of learning to read.

Because of these difficulties, it has been suggested that only one-to-one correspondences be presented to the beginning reader. It is argued that starting with a simplified pattern and later introducing irregularities will make the child understand more readily the notion that English is basically an alphabetic language. Moreover, such a training procedure would presumably also lead to more efficient mastery, in the long run, of all the correspondences that are to be found in the language. A great many of the new approaches to reading instruction subscribe to this idea, for example, the linguistic methods of Bloomfield (1942) and of Fries (1963) and the "initial teaching alphabet" approach (Downing, 1963).

On the other hand, there is also some justification for taking the position that at least some multiple correspondences should be presented right from the start of instruction. The child must indeed learn that the orthography relates to the spoken language, but he must also learn that there are alternative spellings for most phonemes and alternative phonemes for many spellings. Indeed, Levin and Watson (1963) have suggested that a child who is presented with multiple correspondences in the initial stages of instruction will be more likely to develop a useful problem-solving approach to reading (i.e., a "set for diversity"). That is, if he is aware that there is more than one phoneme associated with a particular grapheme, he will be likely to try out a variety of pronunciations when he is faced with an unfamiliar word. In addition to the possibility that an

[1] This research was supported in part by the Cooperative Research Program, United States Office of Education, and in part by a faculty research grant from the University of Pennsylvania. A preliminary report of some of these data was presented at the February 1965 meeting of the American Educational Research Association. The author is indebted to Ellen I. Levy for her assistance in collecting and analyzing the data.

Copyright © 1968 by the American Psychological Association, Inc.

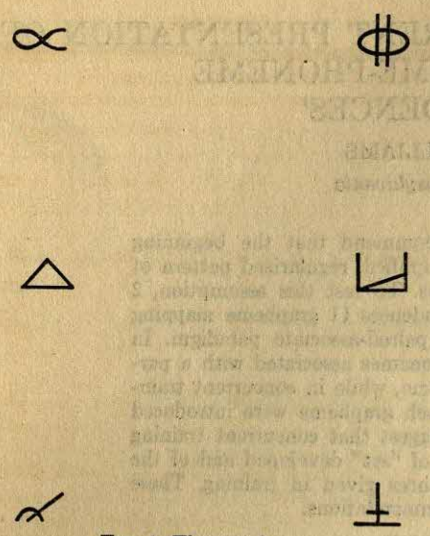


FIG. 1. The graphemes.

effective "set" would be shaped through this kind of training, it is also possible that the correspondences themselves might be learned more effectively in this manner.

In the only data relevant to this question, Levin and Watson (1963) have shown that third-grade children learned a transfer list of words more readily after training on variable correspondences, as compared to training where the correspondences were constant.

When instructional methods are studied in terms of simple laboratory paradigms, it seems especially important to consider not only the effectiveness of training but also its efficiency. A low error-rate during training or superior performance on a test may mean very little in terms of application to actual instruction, if the extra gain has been due to a greatly extended training time. In Levin and Watson's experiments, the original training was continued until Ss reached a performance criterion. The variable-training group took considerably longer to reach criterion, and perhaps the differences in amount of training could account for the differences seen in the transfer task.

The present experiment was designed to compare two methods of teaching multiple correspondences when both methods were given an equivalent amount of training: (*a*) successive: only one of the two phonemes associated with a particular grapheme was presented at a time, and (*b*) concurrent: both phonemes associated with each grapheme were introduced and practiced concurrently. Single (one-to-one) correspondences were also incorporated into the learning task in order to simulate more closely the varied nature of the correspondences found in the actual reading situation.

EXPERIMENT I

Method

Materials. Each of six graphemes, similar to those used by Gibson et al. (1962), was printed on a white 5 × 8 inch card. These graphemes are presented in Figure 1. Ten phonemes were selected, on the basis of pretesting, which were easily produced by *S* and readily discriminable to *E*. These phonemes were: /m/, /a/, /v/, /iy/, /s/, /b/, /uw/, /t/, /k/, /č/. Two of the six graphemes, each of which had two phonemes paired with it, were successive items. That is, one phoneme only was presented during the first half of the training session, and the second phoneme to be associated with that grapheme was presented during the second half of training. Two other graphemes, each of which also had two phonemes paired with it, were concurrent items: both phonemes associated with each grapheme were presented on each trial throughout the entire training period. Each of the last two graphemes had only one phoneme paired with it; these single, one-to-one correspondences were also presented throughout training. The particular graphemes and phonemes used for each type of presentation—successive, concurrent, or single—were balanced over *Ss*, as was the pairing of the graphemes and phonemes.

Subjects. The *Ss* were 36 low-tract fifth and sixth graders. They were asked to volunteer to try out new materials which were being developed to help children learn to read.

Procedure. The *Ss* were shown a sample grapheme (not one used in the experiment) and told that they were to learn the sounds that went with such forms. They were instructed that some forms had more than one sound and that new sounds for some of the forms might be introduced later in training.

A modified paired-associate method was used. Each grapheme was presented individually in a random order on each trial. On the first trial, *S* was simply shown each item in turn. As each grapheme was exposed, *E* pronounced the phoneme or phonemes associated with it, and *S* was asked to repeat the sounds. Each grapheme-phoneme correspondence was given equal training time. That is, during the 10-second exposure of a concurrent item, both phonemes were given once each, and during the 10-second exposure of a successive

item, only the first of its two phonemes was given, and it was given twice. The one phoneme attached to each simple item was given once, and the total exposure was half as long (5 seconds) for each simple item as that of the other types of items.

After the initial trial there were three anticipation trials. Each grapheme was presented for approximately 10 seconds, and *S* was required to give the sound or sounds associated with it. During this part of training, both phonemes were presented for each concurrent item, and only the first of the two phonemes to be associated with each successive item was given.

The second half of training consisted again of one initial trial in which *E* pronounced the phonemes as the graphemes were exposed, and then three anticipation trials. There was no break between the first and second half of training. In this part of the training, the second phoneme associated with each successive item was presented; the other items, of course, remained the same. The *S*s were told after the presentation of each grapheme the correct response for that item. Verbal reinforcement, such as "very good," was given for items on which *S* responded correctly, and *E* drew a star on the data sheet beside every correct answer.

Results

Three tests were given immediately after training. Half of the *S*s received the tests in 1-2-3 order; the other half, in 1-3-2 order. In Test 1, *S*s were simply shown each grapheme individually, and they were asked, "How many sounds did you learn for this form?"

Table 1 presents the number of correct responses as a function of item type and test order. Analysis of variance indicated a significant difference among item types ($F = 17.56$, $df = 2/68$, $p < .001$). There was no difference, of course, as a function of order of presentation of the tests ($F = 2.78$, $df = 1/34$), nor was there any interaction ($F = 2.38$, $df = 2/68$). Specific comparisons showed that *S*s performed equally well on the simple items as on the concurrent items ($F < 1$, $df = 1/68$), and that scores on the successive items were significantly lower than on the other types (successive versus simple: $F = 28.76$, $df = 1/68$, $p < .001$; successive versus concurrent: $F = 11.68$, $df = 1/68$, $p < .005$).

In Test 2, each grapheme was presented, and *E* said, "Give me all the sounds you learned for this form." These data are also presented in Table 1. Analysis of variance was done on Tests 2 and 3 on the two types of multiple correspondences only, because the design of the experiment made it inappropriate to include the simple items in these analyses. Performance was, as on Test 1, better on concurrent items than on successive items ($F = 21.95$, $df = 1/34$, $p < .001$). There was no test-order effect, nor was there any interaction (both *F*s less than 1; $df = 1/34$).

TABLE 1

NUMBER OF CORRECT RESPONSES TO EACH TYPE OF ITEM IN EXPERIMENT I

Test	Type of item (training mode)		
	Single	Successive	Concurrent
Test 1			
Sequence a	32	19	27
Sequence b	29	10	31
Test 2			
Sequence a	25[a]	30	46
Sequence b	21[a]	24	43
Test 3			
Sequence a	28[a]	27	35
Sequence b	26[a]	37	28

Note.—Sequence a = Test 1, Test 2, Test 3; Sequence b = Test 1, Test 3, Test 2.

[a] Number of opportunities equals half of the number for successive and concurrent items.

In Test 3, all the graphemes were presented together in an array. The *E* pronounced each phoneme, and *S* was asked to point to the form that had been associated with that phoneme. After each item, the arrangement of the graphemes in front of *S* was changed (randomly). On this test, performance on the concurrent items did not differ between the concurrent and the successive items ($F < 1$, $df = 1/34$). As on the other tests, there was no effect of test order ($F < 1$, $df = 1/34$), nor was there an interaction between the two variables ($F = 4.08$, $df = 1/34$).

EXPERIMENT II

As described above, the exposure time for the different types of items was equated in this experiment. However, the number of separate presentations of each consecutive correspondence was only half the number of presentations given to each concurrent correspondence. In order to

TABLE 2

NUMBER OF CORRECT RESPONSES TO EACH TYPE OF ITEM IN EXPERIMENT II

Test	Type of item		
	Simple	Successive	Concurrent
Test 1			
Sequence a	32	17	29
Sequence b	30	10	29
Test 2			
Sequence a	27[a]	37	31
Sequence b	24[a]	36	43
Test 3			
Sequence a	25[a]	37	35
Sequence b	28[a]	39	34

Note.—Sequence a = Test 1, Test 2, Test 3; Sequence b = Test 1, Test 3, Test 2.

[a] Number of opportunities equals half of the number for successive and concurrent items.

determine the importance of number of presentations, a second experiment was run in which, on each trial, there were two separate presentations for each successive item. This was not a control; rather, if the original experiment were considered to be biased in favor of concurrent items because of the greater number of presentations, then this experiment was biased in favor of the successive items. While the number of presentations was now constant, the time between any two presentations of the same successive correspondence was, of course, much shorter. Thirty-six *S*s were run, and the data are presented in Table 2.

The overall level of performance was not different from that of the initial experiment (Test 1: $t = .134$; Test 2: $t = .501$; Test 3: $t = .952$, $df = 70$). Again, on Test 1, analysis of variance indicated a difference among item types ($F = 10.20$, $df = 2/68$, $p < .001$), no difference between test orders ($F = 2.68$, $df = 1/34$), and no interaction ($F < 1$, $df = 2/68$). Specific comparisons indicated that performance on simple items was significantly higher than that on successive items ($F = 30.93$, $df = 1/68$, $p < .001$), but did not differ from that of concurrent items ($F < 1$, $df = 1/68$). Performance on concurrent items was significantly higher than that on successive items ($F = 24.27$, $df = 1/68$, $p < .001$).

On Test 2, there was no difference between successive and concurrent items ($F < 1$, $df = 1/34$), contrary to the results of Experiment I. There was no effect of test order ($F < 1$, $df = 1/34$), nor was there an interaction ($F = 2.41$, $df = 1/34$). On Test 3, there were no differences between the two types of items between the test orders, nor was there an interaction (all Fs < 1, $df = 1/34$).

EXPERIMENT III

Further analysis of the original data (Experiment I) showed that many more of the consecutive correspondences that had been presented during the second half of training were given correctly than those presented during the first half (Test 2: $t = 2.707$, $df = 35$, $p < .02$; Test 3: $t = 3.734$, $df = 35$, $p < .01$). That is, there was a recency effect. In order to equate the strengths of the two successive correspondences at the end of training, Experiment III was run, again using 36 *S*s. Here, the six anticipation trials were divided differently. Instead of three trials on the first set of correspondences and then another three on the second set, four trials were given on the first set and two on the second. Again, the overall level of performance (data presented in Table 3) was not different from that of the initial experiment (Test 1: $t = .559$; Test 2: $t = .972$; Test 3:

TABLE 3

NUMBER OF CORRECT RESPONSES TO EACH TYPE OF ITEM IN EXPERIMENT III.

Test	Type of item		
	Simple	Successive	Concurrent
Test 1			
Sequence a	33	17	29
Sequence b	31	8	34
Test 2			
Sequence a	28[a]	42	38
Sequence b	21[a]	32	46
Test 3			
Sequence a	31[a]	44	39
Sequence b	27[a]	37	34

Note.—Sequence a = Test 1, Test 2, Test 3; Sequence b = Test 1, Test 3, Test 2.

[a] Number of opportunities equals half of the number for successive and concurrent items.

$t = 1.720$, $df = 70$). The modification in design was effective in equating the strengths of the two successive correspondences, for there was no difference in the number of times the correspondences were given in the first half of training and the number of times they were given in the second half of training (Test 2: $t = 1.708$, $df = 35$; Test 3: $t = 1.484$, $df = 35$).

As in Experiment II, only Test 1 showed concurrent items significantly superior to successive items (overall $F = 29.98$, $df = 2/68$, $p < .001$; specific comparison $F = 43.79$, $df = 1/68$, $p < .001$) and similar to simple items ($F < 1$) which were significantly superior to successive items ($F = 46.12$, $df = 2/68$, $p < .001$). There were no differences as a function of test order ($F = 1.35$, $df = 1/34$), nor was there an interaction ($F = 2.97$, $df = 2/68$).

On Test 2, neither main effect was significant (type of item: $F < 1$, $df = 1/34$; test order: $F < 1$, $df = 1/34$) and there was no interaction ($F = 3.72$, $df = 1/34$). On Test 3, item type was not a significant effect ($F < 1$, $df = 1/34$), nor was test order ($F < 1$, $df = 1/34$). There was no interaction ($F < 1$, $df = 1/34$).

DISCUSSION

It was suggested above that a comparison of these two methods of training multiple correspondences should be made in terms of (*a*) the level of performance on the correspondences as a function of type of training, and (*b*) the kind of set developed by each type of training. With respect to the latter question, a consideration of Test 1 is relevant, in which *S* was asked how many sounds went with each form. Presumably, if a child does not know that variation is possible, he will not try out several phonemes when he is attacking a new word, and thus he will be less likely to succeed in reading that word. The data indicate that *S*s were better able to identify a grapheme as corresponding to more than one phoneme if the correspondences had been trained concurrently. This suggests that in attempting to read new words, *S*s would more readily identify such graphemes as "multiple" and so try out more than one phoneme, thereby making it more likely that they would read the word successfully.

Further support for the idea that concurrent training fosters the development of a useful problem-solving set comes from the results of an analysis of the errors made on Test 2. Errors were divided into two types: (*a*) incorrect responses, and (*b*) omissions. In Experiment I, the mean proportion of omissions on successive items was .40, whereas on concurrent items this figure was .13. This difference was significant according to a sign test ($z = 2.00$, $p < .05$). The same pattern was seen in the other experiments: Experiment II: successive items: .43; concurrent items: .09; $z = 2.00$, $p < .05$; Experiment III: successive items: .42; concurrent items: .12; $z = 2.33$, $p < .05$.

These data suggest that when *S* knew that an item had two phonemes, as he did on concurrent items, he was more likely to attempt to give two phonemes, guessing when he was not sure.

The second question is whether or not there is a difference between the two training procedures in terms of a simple performance criterion. That is, given equal training time, how many correspondences of each type were learned? Significant differences did in fact appear in favor of the concurrent items on Test 2 (in which *S* was to give all the sounds he had learned for each form). However, this effect was small. It was also quite specific to the particular training conditions, for on Test 3 there were no differences between successive and concurrent training. Moreover, the superiority of concurrent items on Test 2 did not appear in Experiments II and III. Thus, in terms of the performance criterion, there was in some sense a tendency for concurrent presentation to be more effective, but the results were far from clear.

It should be noted, however, that even when extra training and bias in favor of successive presentation was introduced (in Experiments II and III), scores on successive items equalled but never significantly surpassed those on concurrent

items. Thus there seems to be no justification in the present data for the use of successive training methods.

In attempting to make generalizations from this type of laboratory analogue to actual instructional situations, one must keep in mind the differences between the classroom and the experimental situation. One of the more important in the present study, as in Bishop (1964) and Levin and Watson (1963), is that *S*s were older children who had already learned to read. Would a child who is truly naïve about reading and the basic notion of correspondence perform similarly?

To check this, first graders were tested early in their first term of reading instruction. In order to increase the generality of the finding, and also because in many instances the basic unit in instruction is the whole word, the items were homographs, that is, words which have two distinct pronunciations and meanings but only one spelling, such as "wind". Nineteen *S*s were run. A four-item list was used, in which two words were presented as concurrent items and two as successive items. There were no single items. Six trials were run, structured as in Experiment I. Only Tests 1 and 2 were administered. On Test 1, the mean number correct for the successive items was .74, and for the concurrent items, 1.58. This difference was significant at the .01 level ($t = 3.62$, $df = 18$). Performance on successive items (mean number correct = 2.53) was also significantly below that on concurrent items (mean number correct = 3.21) on Test 2 ($t = 3.16$, $df = 18$, $p < .01$). Thus as expected, the first graders' results were similar to those of the main experiments.

Of course, much more research is necessary in order to apply such findings as these to actual instruction. For one thing, few programs teach letter-sound correspondences in isolation. Further work is in progress, focusing on spelling patterns presented in the context of words (as, for example, presented in Fries', 1963, reading materials). At present, however, in the absense of sufficient data on which to base a final decision, it would seem reasonable to provide at least some variation—some kind of concurrent training—when presenting multiple grapheme-phoneme correspondences.

REFERENCES

BISHOP, C. H. Transfer effects of word and letter training. *Journal of Verbal Learning and Verbal Behavior*, 1964, **3**, 215–221.

BLOOMFIELD, L. Linguistics and reading. *Elementary English Revised*, 1942, **19**, 125–30, 183–86.

DOWNING, J. Teaching reading with i. t. a. in Brittain. *Phi Delta Kappan*, 1964, **45**, 322–329.

FRIES, C. C. *Linguistics and reading*. New York: Holt, Rinehart & Winston, 1963.

GIBSON, E. J., GIBSON, J. J., PICK, A. D., & OSSER, H. A developmental study of the discrimination of letter-like forms. *Journal of Comparative and Physiological Psychology*, 1962, **55**, 897–906.

GIBSON, E. J., PICK, H. D., OSSER, H., & HAMMOND, M. The role of grapheme-phoneme correspondence in the perception of words. *American Journal of Psychology*, 1962, **75**, 554–570.

HALL, R. A., JR. *Sound and spelling in English*. Philadelphia: Chilton Books, 1961.

LEVIN, H., & WATSON, J. The learning of variable grapheme-to-phoneme correspondence: Variations in the initial consonant position. In, *A basic research program on reading*. United States Office of Education Cooperative Research Project No. 639, Cornell University, 1963.

WEIR, R., & VENEZKY, R. L. *Rules to aid in the teaching of reading*. United States Office of Education Cooperative Research Project No. 2584, Stanford University, 1965.

(Early publication received January 27, 1968)

Journal of Educational Psychology
1968, Vol. 59, No. 5, 315–319

MANIPULATING THE EFFECTIVENESS OF A SELF-INSTRUCTIONAL PROGRAM

JAMES C. MOORE
University of New Mexico

3 principles of programmed instruction defined as (*a*) gap, (*b*) irrelevancies, and (*c*) mastery were systematically varied in 8 versions of a self-instructional program on test-taking strategies. The criterion of interest was the number of guessing responses an *S* made under each of 4 test conditions. When content identical or similar to questions on the criterion tasks was removed from the instructional materials to create a gap, a significant decrement in performance occurred. When irrelevant material was added to the instructional content and was required to be mastered, a decrement in performance resulted. However, material containing irrelevant instruction which was not required to be mastered did not result in performance decrements. Mastery also interacted with the gap effect.

Although programmed instruction has attracted a good deal of attention, the research directed toward the development of programming principles has been slight. A number of publications (Glaser, 1964; Klaus, 1961; Mager, 1962; Stolurow, 1961; Walther & Crowder, 1965) have discussed and, in part, have demonstrated "principles of programming." Few experimental studies have been reported that have attempted to evaluate the effect on performance of the absence or presence of programming principles singularly or in combination by manipulating them within the content of self-instructional material.

The present study systematically varied three defined classes of variables in a self-instructional program. It was hypothesized that varying program variables in a controlled manner would provide data relating the contribution of these variables to program effectiveness. Conceptually, the research rationale employed was quite similar to that suggested by McClelland (1965). He proposed that once a number of factors had been identified as important in producing a substantial effect with all factors working together to produce it, each factor should then be studied alone to determine its single effect. After a substantial effect has been demonstrated with multiple factors working to produce it, that part of the treatment which deals with each of the factors would be subtracted to discover if there is a significant decline in effect. It should also be possible to omit several factors in various combinations to study interaction effects. The present study applied this conceptual framework to research on principles identified as important in self-instruction.

Principles to be Investigated

Silberman, Coulson, Melaragno, and Newmark (1964) studied exploratory research and individual tutoring techniques with the expressed objective of discovering empirical programming methods which would contribute to the theory of programming. Four self-instructional programs were intensively studied and revised using tutoring techniques with individual students to create modified programs significantly superior to the original ones. Operations performed on the different programs which led to improved student performance were then compared to isolated operations common to all four programs and, therefore, likely to be common to the design of any instructional material. Three principles were isolated: (*a*) gap, (*b*) irrelevancies, and (*c*) mastery. The principles refer, basically, to the addition, elimination, and repetition of instructional material and thus may be identified with specific structural features of the material.

Although these principles may be intuitively obvious on an a priori basis, they bear the added strength of having been induced empirically and can be experimentally manipulated to determine

how their presence or absence within instructional material affects learning performance.

METHOD

Self-Instructional Materials

The basic program used in the study was the Test-Taking Strategy program (Moore, Schutz, & Baker, 1966). The program was designed to develop optimal test-taking strategy. The intrinsic programming technique was used as the instructional method (Crowder, 1959). Specifically, the program deals with how to respond to test questions under the following conditions: (*a*) time limit with no penalty for guessing, (*b*) time limit with penalty for guessing, (*c*) no time limit with penalty for guessing, and (*d*) no time limit with no penalty for guessing. The program explains the operation of correction formulas and describes their effect on test scores. Nine instructional units associated with the following headings are contained in the program:

A. Introduction to scramble book and subject matter
B. Tests with no correction for guessing
C. Tests with correction for guessing
D. Efficient use of testing time
E. Rationale for scoring formulas
F. Time limits
G. Simulated directions and items under a variety of conditions
H. Review of rules
I. Simulated tests

The readability of the program is approximately the fifth-grade level as analyzed by using the Thorndike and Lorge (1944) index. In total, the text is 78 pages in length. By responding to all options correctly with no remedial branching, 39 of the 78 pages would be read. If incorrect responses are chosen to all options, 61 pages would be read. Consequently, most students read between 40–60 pages.

Design and Hypotheses

A 2 × 2 × 2 factorial design (Winer, 1962, pp. 228–258) was used to study each of the three principles: (*a*) gap, (*b*) irrelevancies, and (*c*) mastery; and to test the hypotheses generated. Each of the principles was a main effect with two levels. Five 2 × 2 × 2 analyses were conducted, one for each of five criterion scores. Hypotheses tested were specific to the design. That is, in each analysis, three hypotheses were tested for the three main effects. The three main effects generated four interaction hypotheses. Hypotheses were tested for significance at the .05 level.

Statement of Principles and Manipulation of Program Content

Gap principle. This principle refers to explicit inclusion in the program of instructional units insuring that all criterion skills are covered. The gap principle asserts that criterion performance is increased by the absence of gaps and decreased by the presence of gaps. Two types of frames[1] appear to fill gaps in a program.

1. Frames logically related to the criterion task, which a content analysis indicates are intermediate steps to learning the criterion task.

2. Terminal frames identical or similar to questions on the criterion test, that is, requiring similar responses, or containing similar item content, or both.

The principle was studied by holding frames associated with Type 1 constant while removing frames associated with Type 2. Thus, the discriminative cues for planning a test-taking strategy were provided for (Type 1 frames), but evidence of having learned the skill by application (Type 2 frames) was not required. The remaining instructional material, however, covered the subject matter logically related to the criterion task. Therefore, only the opportunity for application of the concepts covered was removed.

Irrelevancies principle. Irrelevancies refer to unnecessary or distracting frames present in a program. The irrelevancies principle asserts that the presence of irrelevant material will decrease criterion performance. Two types of irrelevant material were studied.

1. Frames not contributing to criterion performance even though they might possess face validity as necessary steps in reaching criterion skills.

2. Frames not at all similar to the criterion task in the nature of the responses required, or in the item content, or both.

Irrelevant frames are analogous to the two types of items that fill gaps, but provide instruction for tasks that are not objectives of the program. While irrelevant frames might possess face validity, a task analysis would indicate that the frames are not necessary for the attainment of the desired instructional outcomes.

Two kinds of irrelevant information were introduced to study the effect of the irrelevancies principle. One type of irrelevant information was instruction on subtasks which were judged to be irrelevant to the terminal objectives of the test-taking strategy program. The subtasks, however, were related to the subject of testing. One subtask concerned the differences between teacher-made and standard tests while the other subtask concerned the differences between objective, completion, and essay tests. The other kind of irrelevant material was the introduction of frames pertaining to correction formulas. While it was judged necessary to provide instruction on the general

[1] For the experimental materials, "frame" was defined as a complete written passage which was informational, or which required the learner to make a response by choosing one of several multiple-choice answers after reading the frame. The majority of frames were 50–150 words in length.

concept of correction formulas as being logically related to the criterion task (Type 1 frame under the gap principle), it was judged irrelevant to require the learner to know the details underlying the concept or to apply the formulas to any data. Thus, in addition to irrelevant instruction on the two subtasks dealing with kinds of tests, this treatment included a large irrelevant proportion of instruction on scoring formulas.

Mastery principle. This principle refers to the provision for mastery of each instructional unit within the program by each learner. Mastery should increase criterion performance while nonmastery should decrease performance. The mastery principle was studied by removing frames associated with alternate amounts of practice as well as all remedial branches. This treatment was analogous to a sequential text in that the learner was not required to demonstrate mastery of each concept before proceeding in the program.

Subjects

The *S*s were eighth-grade students. This particular grade and age group was selected because they had had experience in taking standardized tests as part of the regular school curriculum, but had had no formal instruction in the subject matter presented in the programmed materials used. Since the materials were written at approximately the fifth-grade reading level, data were excluded for those *S*s whose reading performance was below the fifth-grade reading equivalent as measured by the Paragraph Meaning Test of the Stanford Achievement Test, Form JM. This test, with the remainder of the Stanford Achievement Battery, was administered as part of the school's regular testing program 3 months prior to data collection for the study. Thus, the sample was homogeneous to the extent that all *S*s were at the same grade level, of an age range typical to the grade, and evidenced reading performance at the fifth-grade level or higher.

A total of 184 students served in the study which resulted in 23 *S*s under each of the eight treatment conditions prescribed by the design. The number of *S*s was determined by a technique suggested by Federer (1955, pp. 73–76) which was used to compute the approximate replications necessary to assure adequate statistical power. Essentially the technique allows *E* to estimate the number of *S*s required to achieve a desired level of probability for avoiding a Type II error. Power was computed for .95 which indicated a minimum of 20 replicates necessary per treatment condition.

Criterion Instruments

The criterion of interest was the learner's number of guesses under each of four kinds of test conditions: (*a*) time limit with no penalty for guessing, (*b*) time limit with penalty for guessing, (*c*) no time limit with penalty for guessing, and (*d*) no time limit with no penalty for guessing. In order to measure these criteria, four general-information tests of 30 questions each were constructed. Ten of the questions on each test were of moderate to low difficulty for the sample group. Five questions were of such high difficulty that it was extremely unlikely that *S*s could respond except by guessing. Fifteen questions were nonsense questions which had face validity, but any response to them was inferred to be a guessing response. An example of a nonsense question was:

Alaphite mining is an important industry in
a. Brazil b. Panama c. Chile

The four criterion measures were randomly ordered which resulted in the following order of administration:

Test A—no time limit, penalty for guessing.
Test B—no time limit, no penalty for guessing.
Test C—time limit, no penalty for guessing.
Test D—time limit, penalty for guessing.

The content of each criterion test was similar with only the directions regarding penalty and time contingencies varied. Time limits were assigned as a result of pilot studies. The data for the study consisted, therefore, of the guessing scores on the four criterion tests for each *S*. Guessing scores were tabulated by counting the number of high-difficulty and nonsense questions responded to on each test. This resulted in four guessing scores for each *S*. Since there were 5 high-difficulty and 15 nonsense questions on each test, the highest guessing score possible was 20. The remaining 10 questions of low to moderate difficulty were not tabulated. Table 1 reviews the penalty and time characteristics for each test in addition to indicating what the optimal guessing strategy would be under each condition.

TABLE 1
PENALTY AND TIME CONDITIONS OF THE FOUR CRITERION MEASURES AND DESIRABLE GUESSING STRATEGY

Test	Penalty for guessing	Power test	Speed test	High guessing	Low guessing
A	Yes	Yes	No	No	Yes
B	No	Yes	No	Yes	No
C	No	No	Yes	Yes	No
D	Yes	No	Yes	No	Yes

It may be observed from Table 1 that optimal guessing scores would be little or no guessing on Tests A and D because guessing was penalized, and high or total guessing scores on Tests B and C because of no penalty for guessing.

Since the criterion of interest was the number of guesses under specific test conditions on the four tests, it was possible to determine an overall total "appropriateness" score for each individual. This score was computed by giving a point for each nonsense and high-difficulty question not answered on Tests A and D since there was a penalty for guessing, while on Tests B and C a

TABLE 2

DESCRIPTIVE STATISTICS FOR CRITERION TESTS A, B, C, D, AND TOTAL BY TREATMENT CELL FOR EACH OF THE FIVE ANALYSES

Test	M_{Yes}				M_{No}			
	I_{No}		I_{Yes}		I_{No}		I_{Yes}	
	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*
G_{No}								
A	5.17	5.91	10.13	6.05	9.04	7.66	10.87	6.77
B	19.00	3.86	19.35	1.97	17.34	5.55	17.39	5.51
C	17.91	4.06	16.83	4.37	16.70	4.59	17.57	3.46
D	5.83	6.65	8.74	6.75	8.57	7.19	10.44	7.52
Total	65.48	13.39	57.48	12.40	56.43	17.68	53.49	14.04
G_{Yes}								
A	9.30	7.90	12.65	6.16	13.26	6.63	10.70	7.24
B	14.78	6.71	16.87	5.89	17.04	4.32	17.70	4.62
C	12.87	7.37	14.48	6.16	14.30	5.63	15.78	5.56
D	6.83	6.75	11.26	7.89	10.26	6.54	8.17	6.70
Total	51.52	14.82	47.43	15.11	48.26	10.66	52.00	13.14

Note.—A low guessing score was desirable on Tests A and D because of the penalty for guessing condition, while a high score was desirable on Tests B and C because of no penalty for guessing. A high total score was desirable. Abbreviated: M = mastery, I = irrelevancies, G = gap.

point was given for each nonsense and high-difficulty question answered because there was no penalty for guessing. Therefore, the higher an individual's appropriateness score, the better was his total guessing performance for all four tests combined.

Procedure

Prior to the collection of data, *E* consulted with the participating teachers. They were instructed in their procedural responsibilities in administering the materials and in answering questions relevant to the nature of the study.

Data were collected the following week. The experimental programs were arranged in sequential sets of eight, that is, each set had one booklet for each treatment. When the materials were administered, the teacher distributed the booklets by starting with the row of students on his right, giving out the booklets by sets from right to left around the room. Each treatment condition was present in every room. Time was allotted for all *Ss* to complete their material. The criterion instruments were administered 2 days later. Standard directions were used by the teachers in administering both the experimental programs and the criterion tests.

RESULTS

Table 2 presents the descriptive statistics by treatment cell for each of the five $2 \times 2 \times 2$ analyses of variance. The critical F value for all tests at the .05 level of significance was $F = 3.90$, $df = 1/76$.

The F value associated with the gap main effect was statistically significant in four of the five analyses (for Test A, $F = 6.75$; for B, $F = 4.94$; for C, $F = 13.17$; and for Total, $F = 15.80$). Neither the irrelevancies effect nor the mastery effect was significant in any of the five analyses. Two of the interactions proved significant beyond the .05 level, Irrelevancies $\times$ Mastery (for Test A, $F = 4.85$) and Gap $\times$ Mastery (for Test B, $F = 4.94$). The nature of the Irrelevancies $\times$ Mastery interaction may be observed in Table 3 by inspecting the means of the cells associated with the interaction. The analysis of variance for the simple effects of the two levels of mastery for each level of irrelevancies yielded an F of 8.14 ($p < .05$).

Table 4 shows the means associated with the Gap $\times$ Mastery. The analysis for simple effects of the two levels of mastery

TABLE 3

IRRELEVANCIES $\times$ MASTERY CELL MEANS FOR TEST A ANALYSIS

Mastery	Irrelevancies	
	No	Yes
Yes	7.24	11.39
No	11.15	10.78

TABLE 4
GAP × MASTERY CELL MEANS FOR TEST B ANALYSIS

Mastery	Gap	
	No	Yes
Yes	19.17	15.83
No	17.37	17.37

for each level of gap resulted in an F of 9.88 ($p < .05$).

DISCUSSION

The most predictable of the findings was the effect produced by introducing a gap into the instructional material. The most interesting results, however, are those associated with the interaction effects of the mastery and irrelevancies principles. When material irrelevant to the objectives was introduced and mastery of it was required, Ss employed significantly poorer guessing strategies. On the other hand, irrelevant material, when presented, but not required to be mastered, did not produce a detrimental effect on performance. The following tentative conclusions appear appropriate:

1. If the task to be learned can be simulated, instructional materials should include tasks identical to or highly similar to the criterion task as part of the program. Although learning appears to result from instruction which does not include simulation of the criterion task, it occurs to a lesser degree (gap principle).

2. If information is included in the instructional material which a task analysis indicates is not essential to reaching the terminal objectives, mastery of the nonessential material should not be required. However, if irrelevant material with face validity is introduced and is not required to be mastered, there is no apparent decremental effect on learning (Irrelevancies × Mastery interaction).

3. If gaps are not introduced into the instructional material, mastery of the material should be required. However, if gaps are introduced, mastery of the remaining material does not appear to benefit instruction (Gap × Mastery interaction).

4. The principle of mastery of instructional material should be used selectively to require mastery only of that instruction which a task analysis indicates is essential to obtaining the terminal objectives.

REFERENCES

CROWDER, N. A. Automatic tutoring by means of intrinsic programming. In E. H. Galanter (Ed.), *Automatic teaching*. New York: Wiley, 1959.

FEDERER, W. T. *Experimental design*. New York: MacMillan, 1955.

GLASER, R. Programmed instruction: A behavioral view. In A. I. Grazia & D. A. Sohn (Eds.), *Programs, teacher and machines*. New York: Metron, 1964.

KLAUS, D. J. The art of auto-instructional programming. *Audiovisual Communication Review*, 1961, **9**, 130–142.

MCCLELLAND, D. C. Toward a theory of motive acquisition. *American Psychologist*, 1965, **20**, 321–333.

MAGER, R. F. *Preparing objectives for programmed instruction*. San Francisco: Fearon Press, 1962.

MOORE, J. C., SCHUTZ, R. E., & BAKER, R. L. The application of a self-instructional technique to develop a test-taking strategy. *American Educational Research Journal*, 1966, **3**, 13–17.

SILBERMAN, H., COULSON, J., MELARAGNO, R., & NEWMARK, G. Use of exploratory research and individual tutoring techniques for the development of programming methods and theory. Final Report. Title VII, National Defense Educational Act grant 7-14-0000-181, 1964.

STOLUROW, L. M. Teaching by machine. *Cooperative Research Monograph*, (OE-34010) Washington, D. C.: U. S. Government Printing Office, 1961.

THORNDIKE, E. L., & LORGE, I. *The teacher's word book of 30,000 words*. New York: Teachers College, Columbia University, 1944.

WALTHER, R. E., & CROWDER, N. *A guide to preparing intrinsically programmed instructional materials*. AMRL-TR-65-43. Wright-Patterson Air Force Base, Ohio: Behavioral Sciences Laboratory, 1965.

WINER, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

(Received August 30, 1966)

Journal of Educational Psychology
1968, Vol. 59, No. 5, 320–324

ATTITUDINAL AND INTELLECTUAL CORRELATES OF ATTENTION: A STUDY OF FOUR SIXTH-GRADE CLASSROOMS[1]

HENRIETTE M. LAHADERNE[2]
University of Chicago

Data collected from 4 6th-grade classrooms ($N = 125$) were examined to determine whether children's attentiveness in class was related to their attitudes toward school on the one hand, and to achievement and ability on the other. Each pupil's attention to the main class activity was recorded over a 2-month period, questionnaires assessing the attitudes were administered, and IQ and achievement-test scores were obtained from school records. There was practically no relation between students' attitudes and measures of attention; however, a positive relationship was found between measures of students' attention and scores on achievement and intelligence tests. In sum, all of the pupils in a classroom may have been subjected to the pressures for attention but the extent to which they responded appears tied to a general ability variable rather than to an attitudinal one.

Teachers gauge the success of their teaching not so much by the scores their pupils attain on achievement tests as by the involvement pupils demonstrate during ongoing class activities (Jackson & Belford, 1965). They assume that if a child is engrossed in an activity, he is getting something out of it even if that "something" is not identifiable or measurable. This way of looking at things may seem at odds with recommended procedures of evaluation but it makes sense when considered in the social context of the classroom.

In social gatherings, the individual communicates his esteem and attachment for the other members as well as for the situation itself by giving or withholding his attention to the activity at hand (Goffman, 1963). From this viewpoint, it is not surprising that the teacher, as a leader of a social gathering and responsible for engaging the pupils, should be alert to the behavioral cues of involvement. Moreover, the flow of classroom life cannot wait for a delayed measurement. Small wonder, then, that standard test scores offer less immediate feedback to the teacher than do the flickering signs of attention.

The concern of the present study is whether these fleeting cues are related to more enduring characteristics of students. Is it possible, for example, that attention to specific tasks presages both a positive orientation toward school and academic gains? Perhaps the behaviors that serve as cues of attention are also conducive to the development of satisfaction with school and academic performance. Partial support for such a possibility comes from a study of classroom behavior in an Air Force school. Among the findings was a correlation of −.58 between achievement and student behavior indicating inattention (Morsh, 1956). Pursuing this line of inquiry, the present study examines whether attention is related to attitude toward school and the teacher on the one hand, and to academic achievement and ability, on the other.

METHOD

The classroom behavior of pupils was observed over a 3-month period, questionnaires were administered to the pupils, and such background information as IQ and achievement-test scores was obtained from school records.

---

[1] Expanded version of a paper presented at the American Educational Research Association Convention, New York, New York, February, 1967. The research reported herein was performed pursuant to a contract with the United States Department of Health, Education, and Welfare, Office of Education, under the provisions of the Cooperative Research Program. The author is grateful to Philip W. Jackson for his aid and encouragement.

[2] Now at the Institute for the Development of Educational Activities, Inc.

Subjects

The *Ss* were 125 pupils (62 boys and 63 girls) enrolled in four sixth-grade classrooms located in a predominantly white, working-class suburb. The pupils' mean IQ as measured by the Kuhlmann-Anderson Intelligence Test was 104 for the boys and 110 for the girls. The standard deviations were 14.5 and 13.4 for boys and girls, respectively. Two of the classes, containing 33 and 34 pupils and taught by men, were in one school; the other two, each containing 29 pupils and taught by women, were in another school.

Observations

The visits, which ranged from ½ hour to a full day, began in late September and continued through November. During each visit periodic tallies of pupil attention were made along with other observations which are not relevant to this paper. As far as was possible the observations were distributed over the entire school week and they sampled most of the activities in each room. The total hours of observation was 37, or approximately 9 hours in each of the four classrooms.

Using a modified version of the Jackson-Hudgins Observation Schedule (Jackson & Hudgins, 1965), the observer looked at each pupil in turn and immediately recorded the state of his attention.[3] Four classifications were possible:

1. "+" if the pupil was attentive. The pupil had to be attending to the area of focus, namely, the subject to which the teacher had called attention, for example, arithmetic, social studies, or art. The pupil also had to be attending to the prescribed activity, that is, the activity designated by the teacher, such as writing in an arithmetic workbook or reading in a social studies textbook.

2. "−" if the pupil was clearly inattentive. The pupil was marked inattentive if he were not attending to the area of focus and/or the prescribed activity. This classification included instances of horseplay, reading a book when writing had been prescribed, and doodling when attention should be focused on the blackboard.

3. "?" if it was uncertain to the observer whether or not the pupil was attentive.

4. "0" if the pupil's attention was not observable.

Interobserver reliability, defined as percentages, ranged from 83% to 100% in trial observations. Additional evidence of the reliability of the method has been reported by other investigators. For example, interobserver reliability, defined as percentages of agreement, ranged from 85% to 100%, with a median of 90% for a series of observations (Hudgins, 1967).

[3] A description of the conventions set up for recording attention is included in the author's final report, "Adaptation to School Settings: A Study of Children's Attitudes and Classroom Behavior," Washington, D.C.: Educational Research Information Center, 1967.

Questionnaires

Student Opinion Poll II. The children's attitudes toward school were measured by the Student Opinion Poll II. This is a 47 multiple-choice-item test derived from an original 60-item test (Jackson & Getzels, 1959). The questions concern four aspects of school life, namely, the curriculum, the teacher, the peers, and the school. The following are sample items:

6. The things I am asked to study are of:
   a. great interest to me
   b. average interest to me
   c. little interest to me
   d. no interest to me

47. In general, my feelings toward school are:
   a. very favorable—I like it as it is
   b. somewhat favorable—I would like a few changes
   c. somewhat unfavorable—I would like many changes
   d. very unfavorable—I frequently feel that school is pretty much a waste of time.

The test was scored by giving one point each time the student chose, from a set of multiple-choices, the response indicating the highest degree of satisfaction with that aspect of school life under question. Thus the possible range of scores was from 0 to 47. The mean scores were 28 for the boys and 31 for the girls. The standard deviations were 8.9 and 7.2 for boys and girls, respectively. The test was readministered to 63 pupils after a 5-month interval; the rank correlation coefficient between the pupils' two scores was .66. The coefficient of reliability, based on the Kuder-Richardson formula 20, was .89 for the boys, and .85 for the girls. In an earlier study, involving 293 sixth graders, the test reliability was .86.

The Michigan Student Questionnaire (abbreviated version). An abbreviated version of the Michigan Student Questionnaire (Flanders, 1965) assessed the students' attitude toward their present teacher and schoolwork. The form used in this study contained 37 descriptive statements, each followed by four possible replies: strongly disagree, disagree, agree, and strongly agree. A student's response to each item was scored 1, 2, 3, or 4 depending on the degree to which it revealed a positive attitude toward his teacher. Hence, the possible range of scores was from 37 to 148. The mean scores for the sample of sixth graders were 110 for the boys and 114 for the girls. The standard deviations were 14.3 and 11.2 for boys and girls, respectively. Test reliability based on a variation of the Kuder-Richardson formula appropriate for weighted scores (Ferguson, 1951) was .94 in a study involving 293 sixth graders. The following are sample items:

16. This teacher certainly knows how to teach.

Strongly disagree	Disagree	Agree	Strongly agree

23. I really like this class.
Strongly disagree Disagree Agree Strongly agree

Achievement and IQ

The achievement test scores were derived from: (*a*) the Scott-Foresman Basic Reading Test to accompany *The New People and Progress;* and (*b*) the Stanford Achievement Test (Intermediate II, complete battery). The intelligence quotient was taken from the Kuhlmann-Anderson Intelligence Test.

TABLE 1

CORRELATION BETWEEN ATTENTION AND STUDENTS' ATTITUDES

Attention	Attitudes: Student Opinion Poll II		Attitudes: Michigan Student Questionnaire	
	Boys[a]	Girls[b]	Boys[c]	Girls[d]
Attentive	.12	−.13	.02	−.09
Inattentive	−.07	.10	.00	.03
Uncertain	−.08	.10	−.02	.11
Nonobservable	−.16	.19	−.09	.22

[a] $N = 62$.
[b] $N = 63$.
[c] $N = 61$.
[d] $N = 63$.

RESULTS

The most noteworthy finding in Table 1 is an overall lack of relation between student attitudes and attention. For neither boys nor girls were feelings toward the school and the teacher related to their attention to the dominant class activity.

The correlations in Table 2 between attention and measures of achievement support what seems self-evident—the pupil who paid attention gained the most from his instruction. Or, conversely, the data might be said to show that the pupil who was inattentive was not apt to achieve academically. This finding confirms that of the study cited earlier in which Air Force students' inattention was negatively associated to achievement (Morsh, 1956).

Table 3 shows a relation between attention and IQ.[4] The brighter the pupil, the more he was likely to be attentive in class. This raises the obvious question of whether attention makes a unique contribution to achievement or whether its effect is due solely to its linkage to IQ. A

[4] A check was made to ascertain whether the measure of attention had been contaminated by observer bias. Conceivably, the observer gradually might have acquired knowledge of each pupil's ability and eventually she might have judged each pupil on the basis of his ability rather than his behavior. If such a bias existed, the correlation between attention and IQ would have increased systematically with each successive period of observation. This possibility was tested by plotting IQ scores against attention measures during the first third of the observations, the second, and, finally, the last third of the observations. No systematic bias was apparent.

TABLE 2

CORRELATIONS BETWEEN ATTENTION AND MEASURES OF ACHIEVEMENT

Attention	Scott-Foresman Reading		Stanford Reading		Stanford Arithmetic		Stanford Language	
	Boys[a]	Girls[b]	Boys[c]	Girls[d]	Boys[c]	Girls[d]	Boys[c]	Girls[d]
Attentive	.51**	.49**	.46**	.39**	.53**	.39**	.48**	.37**
Inattentive	−.47**	−.53**	−.42**	−.44**	−.52**	−.39**	−.47**	−.38**
Uncertain	−.28*	−.33**	−.37**	−.24	−.36**	−.37**	−.34**	−.31**
Nonobservable	−.23	.07	−.08	.05	−.06	.17	−.03	.11

[a] $N = 61$.
[b] $N = 63$.
[c] $N = 56$.
[d] $N = 55$.
* $p < .05$.
** $p < .01$.

regression analysis focusing on the partial contribution of each variable to achievement was performed and revealed some evidence of the singular effect of attention on achievement. For boys the partial correlation coefficient between achievement and attention, with IQ held constant, was .31 ($p < .05$) with the Scott-Foresman Reading Test, and .26 ($p < .05$) with the Stanford Arithmetic Achievement Test. For girls, however, a statistically significant result was obtained with only the Scott-Foresman Reading Test. The partial correlation coefficient between that test and attention, with IQ held constant, was .26 ($p < .05$). Apparently, attention makes a difference with respect to certain types of achievement but not others. More important is the question of whether it is proper to search for the effect of attention independent of IQ. Maybe the ability to attend is an integral part of intelligent performance and contributes as much to a child's performance on an IQ test as to his achievement in school.

Finally, Table 4 shows low correlations between students' attitudes and their achievement-test scores and IQ. The lack of a connection between attitude toward school and scholastic performance confirms the findings of prior studies (Jackson & Lahaderne, 1967). Despite the repeated absence of ties between the attitude scores and the measures of scholastic performance and pupil attention, the Student Opinion Poll II and the Michigan Student Questionnaire correlate significantly with each other ($r = .63$; $p < .001$) and with

TABLE 3

CORRELATIONS BETWEEN ATTENTION AND IQ

Attention	IQ	
	Boys[a]	Girls[b]
Attentive	.48**	.44**
Inattentive	−.35**	−.46**
Uncertain	−.49**	−.33**
Nonobservable	−.20	.07

[a] $N = 61$.
[b] $N = 63$.
* $p < .05$.
** $p < .01$.

TABLE 4

CORRELATION BETWEEN STUDENTS' ATTITUDES AND MEASURES OF SCHOLASTIC PERFORMANCE

Measures	Student Opinion Poll II		Michigan Student Questionnaire	
	Boys	Girls	Boys	Girls
Achievement				
Scott-Foresman Reading[a]	.17	.05	.01	−.01
Stanford-Reading[b]	.16	−.10	.08	−.12
Stanford-Arithmetic[b]	.16	.03	.01	.02
Stanford-Language[b]	.07	−.08	−.05	−.07
IQ[a]	.15	.10	.08	−.06

[a] $N = 61$ boys and 63 girls.
[b] $N = 56$ boys and 55 girls.

measures of other variables. The Student Opinion Poll II, for example, correlates significantly with the Children's Intellectual Achievement Responsibility Questionnaire ($N = 292$; $r = .41$; $p < .001$), the Children's Social Desirability Questionnaire ($N = 125$; $r = .33$; $p < .001$), and teachers' ratings of their pupils' attitude toward school ($N = 292$; $r = .35$; $p < .001$).

DISCUSSION

The real problem posed by the results of this study concerns the lack of a connection between the way students felt about school and their attentiveness in class. Why, for example, did students who were dissatisfied with school appear to be just as attentive as those who were satisfied? What has happened to the popular stereotype of the daydreaming malcontent?

Perhaps the constraints imposed on pupils to be attentive were so strong that attitudes could not influence behavior. Consider, for example, the following restrictions. Pupils could not leave the classroom, or for that matter, get up from their desks without permission. They could not chatter with their neighbors. They had to be recognized before speaking up in class. Their actions at any given moment had to be within the sphere prescribed by the teacher. Moreover, one of the teacher's major functions was to preserve the classroom order. She called on the reluctant, snapped the daydreamer back to attention, reprimanded the cutup, and often

reminded the pupils of the designated focus of attention. In short, pupils were coaxed and compelled to adhere to a code of conduct that supported the order of the classroom. Thus, regardless of how he felt about school, the disgruntled pupil had little chance to do anything about it in the classroom.

It is evident that the forces for attention impinged on everyone. Less apparent are the variables that accounted for fluctuations of attention. The data in Table 3 suggested the possibility that ability to attend may be an integral part of intelligent behavior. If this were indeed the case, the less able pupils may have been limited in their capacity to attend just as they were in their capacity to achieve academically. Furthermore, the usual classroom situation where the teacher directed the curriculum to what he considered was the class average may have strengthened the connection between intelligence and attention. The able may have understood and participated in the instructional matter but the less able could not keep up. This possibility implies that a curvilinear relation may exist between attention and level of instruction.

As the level of instruction increases in difficulty from zero, attention may also increase to an optimal point and then decrease beyond that point. Moreover, it seems likely that the optimal point may vary with ability. The apex for the high-ability pupils may be at a higher point than for low-ability pupils. According to this speculation, the less able pupils in the present study may have been inattentive when the level of instruction was beyond their optimal point, and the brighter pupils, when the level of instruction was below their optimal point. In sum, all of the pupils in a classroom may have been subjected to the pressures for attention but the extent to which they responded appears to have been tied to general ability and instructional variables rather than to the pupils' attitude toward school.

REFERENCES

FERGUSON, G. A. A note on the Kuder-Richardson formula. *Educational Psychology Measurement,* 1951, **11**, 612–615.

FLANDERS, N. A. Teacher influence, pupil attitudes, and achievement. Cooperative Research Monograph, 1965, No. 12 (OE-25040), Washington: United States Government Printing Office.

GOFFMAN, E. *Behavior in public places.* New York: Free Press of Glencoe, 1963.

HUDGINS, B. B. Attending and thinking in the classroom. *Psychology in the Schools,* 1967, **4**, 211–216.

JACKSON, P. W., & BELFORD, E. Educational objectives and the joys of teaching. *The School Review,* 1965, **73**, 267–291.

JACKSON, P. W., & GETZELS, J. W. Psychological health and classroom functioning: A study of dissatisfaction with school among adolescents. *Journal of Educational Psychology,* 1959, **50**, 295–300.

JACKSON, P. W., & HUDGINS, B. Observation schedule for recording pupil attention. Unpublished, 1965.

JACKSON, P. W., & LAHADERNE, H. M. Scholastic success and attitude toward school in a population of sixth graders. *Journal of Educational Psychology,* 1967, **58**, 15–18.

MORSH, J. E. Development report—systematic observation of instructor behavior. USAF: *Personality Training Research Center Development.* 1956, No. AFPTRC TN 56-52.

(Received February 24, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 5, 325–327

SEMANTIC SPACE AS AN INDICATOR OF SOCIALIZATION

KEITH A. McNEIL[1]
The University of Texas, Austin

A semantic differential was administered to 521 6th-grade *Ss* who could be classified into 4 subcultural groups representing an hypothesized degree of socialization. A principal components factor analysis yielded 6 factors having an eigenvalue above 1. Factor scores were computed for each *S* and subjected to a discriminant analysis. The discriminant function which accounted for the most variance produced the same rank ordering of the subcultures as the hypothesized degree of socialization. The degree of socialization of an individual can thus be estimated from the empirical weights obtained in the discriminant analysis.

Several investigators have claimed that the structure of semantic space is similar across widely different cultures (Osgood, Suci, & Tannenbaum, 1957; Tanaka, Oyama, & Osgood, 1963). Several problems, including translation equivalence (Brown, 1958) and developing an adequate statistical technique with which to match factors, suggest caution in accepting the above conclusions. Also, only gross structural similarities have been reported, and it is quite feasible that subtle semantic differences may be very important. For example, communication is sometimes hindered when the parties involved do not attach the exact same meaning to the words being used in the communication act.

The well-documented subcultural differences in linguistic performance, as well as in other behavioral realms, suggest that there may indeed be differences in subcultural meaning. Subcultural differences in meaning, if not apparent in the basic structure of semantic meaning, should be apparent at least in the magnitude on each semantic dimension. Evidence bearing on the structural differences of subcultural meaning has been reported elsewhere (McNeil, 1967), whereas evidence pertaining to subcultural differences in magnitude of semantic meaning is presented here.

Concomitant with an increased level of socialization may be a change in one's view of the environment. This change should be evident in one's connotative meaning. Thus, looking at one's connotative meaning may provide information as to the level of socialization. The present investigation is an attempt to predict level of socialization through measurement of connotative meaning. In particular, it was expected that the more socialized a subculture, the more closely that subculture's profile of semantic meaning would resemble the core culture's profile of semantic meaning.

METHOD

Sample

The four subcultural groups in this study communicate and compete in the same culture. The extent to which the various subcultures have succeeded in "attaining the core culture," that is, in being socialized, was hypothesized as follows, from most to least: middle class whites (MCW); lower class whites (LCW); lower class Negroes (LCN); lower class Latin Americans (LCLA). The LCLA were hypothesized to be at the bottom of the socialization continuum because of the foreign language influence in their homes and also because of two cultural aspects. "Mañana" is a fatalistic attitude expressed by taking care of today's needs and not worrying about tomorrow's needs (Saunders, 1954). The LCLA's outlook is thus different from that of the outlook of the core culture. "Evidia" is a type of black magic wherein negative sanctions are applied to those who achieve some degree of success (Madsen, 1964). To the extent that evidia operates in the subculture, the aspirations of the members are restricted.

Instrumentation

A semantic differential was administered by trained graduate students to 521 sixth-grade Austin children. The instrument contained 12 high frequency concepts (Elephant, Army, Me, Butter, Fear, Clouds, Doctor, Street, Baby, Schoolwork, Policeman, Pain), each being rated on 20 bipolar adjectives. The concepts and scales were selected on the basis of previous research, mainly the work reported by Lilly (1965). Each bipolar adjective constituted a 7-point scale. Scale responses were

[1] Now in the Department of Guidance and Educational Psychology at Southern Illinois University.

TABLE 1

VARIMAX LOADINGS OBTAINED FROM THE TOTAL SAMPLE OF 521 SUBJECTS

Semantic differential scales	I	II	III	IV	V	VI
Strong-Weak	74	−08	08	−16	14	−07
Good-Bad	51	−28	20	09	42	04
Usual-Unusual	04	−40	00	51	32	−22
Fast-Slow	69	−07	−05	13	02	09
Interesting-Boring	60	−37	29	12	05	02
Not important-Important	−14	77	11	−01	−12	−15
Uncertain-Certain	−22	64	15	−01	−07	−27
Changing-Steady	09	−04	−06	71	−04	−13
Small-Large	−11	15	78	10	−17	−09
Rough-Smooth	16	23	−32	23	−15	−55
Heavy-Light	38	08	−16	−22	01	−61
Old-New	−17	−04	06	11	17	−72
Moving-Still	60	−13	−16	29	−09	−11
Simple-Complex	20	24	17	50	06	28
Dirty-Clean	−15	38	23	16	−40	−50
Not active-Active	−08	75	11	04	06	04
Soft-Hard	25	14	71	−13	08	25
Near-Far	17	06	−09	08	71	−19
Dangerous-Safe	15	27	01	03	−60	−40
Controlled-Not controlled	12	−56	17	15	36	−07

Note.—Decimal points have been omitted.

summed across the 12 concepts, and correlated to yield a 20 × 20 correlation matrix. The resultant correlation matrix was factored (with unities in the diagonal) by the principal axes method, with Varimax rotation of factors whose eigenvalues exceeded 1.0 (Veldman, 1967). Six factors resulted and factor scores were computed on the six factors for each *S*. We thus have a measure of each *S*'s semantic meaning on each of the important dimensions of semantic meaning.

The *S*s were then grouped into their respective subcultural classifications and a multiple discriminant analysis (Veldman, 1967) was performed on the four subcultures, considering the six factor scores as the variables on which the discrimination was to be made.

RESULTS

Table 1 contains the factor loadings on the six obtained semantic factors. Interpretation of the factors is not crucial to the present study, so just a brief listing of the factor names seems sufficient. Factor I—Evaluation; Factor II—Activity; Factor III—Potency; Factor IV—Stability; Factor V—Security; Factor VI—Unnamed.

TABLE 2

SUBCULTURAL FACTOR SCORE MEANS

Factor	Middle class white $N = 134$	Lower class white $N = 146$	Lower class Negro $N = 83$	Lower class Latin American $N = 158$	F^a	p
Evaluation	.34	.12	−.37	−.20	12.36	<.0001
Activity	.17	.09	.22	−.34	9.64	<.0001
Potency	−.01	.16	−.32	.03	4.13	<.01
Stability	−.13	−.06	.31	.00	3.61	<.02
Security	−.29	−.01	.08	.21	6.38	<.001
Unnamed	.16	.14	.00	−.27	6.04	<.001

Note.—Factor scores based on the factor analysis of the total sample.
[a] *df* within *S*s = 517; *df* between *S*s = 3.

Table 2 indicates that there are significant differences between the factor scores of the four subcultures on each of the six semantic factors. Consistent trends cannot be discerned, nor does an overall interpretation seem to be possible with these univariate analyses.

Table 3 indicates that there are two orthogonal ways of combining the factor scores so as to separate significantly the four subcultures. The first discriminant function (Table 4) is of most interest, though, as it accounts for the most variance. The first function produces group centroids (group means of the discriminant scores) which are in the same relative order as the hypothesized degree of socializa-

TABLE 3

DISCRIMINANT ANALYSIS RESULTS: GROUP CENTROIDS ON THE THREE DISCRIMINANT FUNCTIONS

Subcultures	Discriminant functions 1	2	3
Middle class white	.51	−.02	−.11
Lower class white	.20	−.07	.15
Lower class Negro	−.27	.56	.00
Lower class Latin American	−.48	−.21	−.04
Variance	70.03%	26.05%	3.92%
χ^2	88.43[a]*	34.69[b]*	5.38[c]

[a] *df* = 8.
[b] *df* = 6.
[c] *df* = 4.
* $p < .0001$.

TABLE 4
DISCRIMINANT WEIGHTS FOR THE FIRST FUNCTION

Factor	Weight
Evaluation	.606
Activity	.429
Potency	.108
Stability	−.218
Security	−.454
Unnamed	.428

Note.—Wilk's lambda = .780; $F = 7.407$, $df = 18/1{,}449$, $p < .0001$.

tion. The second discriminant function also produces a significant difference between the four subgroups, and it should be noted that the differences between the two nonwhite subcultures are responsible for the definition of the two ends of the discriminant function. Interestingly, the third discriminant function (nonsignificant) was defined by the two white subcultures.

DISCUSSION

The results of the discriminant analysis indicate that the four subcultures can be separated significantly, when scores on semantic factors are taken into consideration. What is of more importance is the finding that the relative order of the subcultures on the discriminant function which accounts for the most variance is the same as the hypothesized degree of socialization. Applying the obtained discriminant weights (Table 4) to semantic factor scores of a new *S*, would predict his position on the first discriminant function (or estimate his degree of socialization).

After the majority of the variance is extracted from the data by the first discriminant function, the underlying function which extracts the most amount of variance is due to the semantic differences between the two nonwhite subcultures. After these two sources of variance are accounted for, remaining variance is accounted for by the two white subcultures. Thus, the most important underlying dimension accounting for the subcultural differences in semantic meaning is not due to ethnicity or social class, but to the degree of socialization.

The extent to which the obtained results are a consequence of subcultural differences unrelated to socialization needs to be verified. That is, the results need to be cross-validated on a sample of *S*s who can be classified into groups of varying degrees of socialization, on the basis of a socialization variable.

It is further suggested that the semantic differential technique would be a valuable instrument to measure socialization because of the reduction of the effects of response sets. That is, the instrument does not appear to *S* to be measuring socialization. In fact, the average *S* is unable to determine how the instrument is being used. Also, the instrument can yield additional information, besides a socialization score, which could be valuable.

REFERENCES

BROWN, R. Is a boulder sweet or sour? *Contemporary Psychology*, 1958, **3**, 113–115.

LILLY, R. S. A developmental study of the semantic differential. *Educational Testing Service Research Bulletin* (RB-65-28), 1965.

MADSEN, W. *The Mexican-American of South Texas*. New York: Holt, Rinehart & Winston, 1964.

MCNEIL, K. Multivariate relationships between the semantic space of various subcultures and selected personality variables. Unpublished doctoral dissertation, University of Texas at Austin, 1967.

OSGOOD, C., SUCI, G., & TANNENBAUM, P. *The measurement of meaning*. Urbana: University of Illinois Press, 1957.

SAUNDERS, L. *Cultural differences and medical care.* New York: Russell Sage Foundation, 1954.

TANAKA, Y., OYAMA, T., & OSGOOD, C. Cross-cultural and cross-concept study of the generality of semantic space. *Journal of Verbal Learning and Verbal Behavior*, 1963, **2**, 392–405.

VELDMAN, D. *FORTRAN programming for the behavioral sciences*. New York: Holt, Rinehart & Winston, 1967.

(Received June 27, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 5, 328-333

SELECTION OF DEFINING PROPERTIES IN CONCEPT ATTAINMENT

D. CECIL CLARK[1] AND FREDERICK J. McDONALD
Stanford University

The present study explored the question, When an S is presented with several examples of a concept, what determines which of the common properties he will select as defining the concept? 3 competing hypotheses were tested by asking 5th graders to select a defining property when presented with verbal examples of a concept. Results showed that (*a*) the common property with the highest total association strength will be selected as defining with significantly greater frequency than other common properties, and (*b*) neither the order in which the examples are presented nor the presence of an example with a strongly associated property appears to influence the selection of the defining property.

Many of the concepts taught in the elementary classroom are of the type that have clearly differentiated properties which must be accurately learned by all of the students. In teaching for this type of concept, the teacher usually presents several examples, pointing out the defining properties common to each. Thus, in teaching for the concept "tree," she may show a picture of a maple tree, a picture of a cottonwood tree, and a picture of a douglas fir tree, indicating that, among other properties, each has a crown, a trunk, and roots (defining properties, see A, Table 1). In testing to determine whether the class has mastered the concept, she observes that some of the students correctly identify new pictures of trees and nontrees, while others do not. One explanation for this phenomenon is that students differ in the common properties they select to define the concept. One student might correctly select crown, trunk, and roots as defining properties while another student, from the same examples, might incorrectly select color, trunk, and size as defining the concept.

Such differences give rise to a basic question about the process of forming concepts: When presented with several examples of a concept, what determines which of the common properties will be selected as defining the concept? The answer to this question has implications for the optimal type and sequence of examples that can be used to illustrate a given concept.

Background

In a typical concept presentation the teacher may offer both pictorial and verbal examples to illustrate a given concept (A, Table 1). This study was limited to the selection of defining properties when only verbal examples were used (B, Table 1). Thus, properties such as color, size, and shape which are often significant with pictorial examples were not considered in the present study.

Underwood and Richardson (1956a) have developed a set of adult word-association norms. When presented with a noun, *S*s were asked to respond with any sensory adjective that came to mind. Responses to each noun as well as their frequencies of occurrence were tabulated. For example, 61% of the *S*s responded to the noun *button* with the sensory adjective "round," 15% with "small," 5% with "hard," and 5% with "white." According to their norms, the same adjective was often given to different nouns. For instance, "round," "hard," "small," and "white" were likewise common adjectives for *baseball* and *hailstone* (B, Table 1). Associative strength represented the percentage of *S*s giving a particular adjective to a particular noun. Thus, the associative strength between *button* and "round" was 61%, and between *button* and "small," 15%.

[1] Now Associate Professor, Department of Educational Psychology, College of Education, University of Washington.

TABLE 1

TAXONOMY OF A CLASSROOM-TYPE CONCEPT AND AN EXPERIMENTAL CONCEPT

Example	Common properties	Concept
Classroom (A)		
Picture of maple tree[a] Picture of cottonwood tree[a] Picture of douglas fir tree[a]	Crown[b] Trunk[b] Roots[b] Brownish Large branches	Tree
Experimental (B)		
Button Baseball Hailstone	Round Hard Small White	X

[a] Accompanied by explanation.
[b] Arbitrarily selected as defining properties.

In the present study, the nouns *button, baseball,* and *hailstone* were viewed as three examples of some new concept, "X," and "round," "small," "hard," and "white" represented properties common to the three examples (B, Table 1). The basic question, then, was what would cause one of these common properties to be selected over the others as defining the concept, "X"?

HYPOTHESES

Three alternative explanations provided the bases for the hypotheses tested in this study. One explanation was that as *S* thinks about the properties common to the three examples, a combining operation occurs. The three associative strengths of a property are added together; this occurs with the other three properties as well. (In Table 2 for instance, "round" = 61 + 70 + 14 = 145.) The first hypothesis was that the property with the highest total associative strength would be selected as defining concept, "X," with greater frequency than the other properties. (This study was limited to only one defining property.) Thus, in Table 2, since "round" has a higher total associative strength than the other properties, it would be selected in defining concept, "X." Results of Underwood and Richardson (1956a), Schulz, Miller, and Radtke (1962), and Coleman (1964) suggest the plausibility of this hypothesis.

A second competing explanation was that the strength of the association between *one* of the examples and *one* of its properties would be so strong that other properties of this example—and dissimilar properties of other examples—would become less obvious. The effect of this strong association in one example would be that of highlighting the same property in other examples. The second hypothesis was that the common property with the single greatest associative strength would be selected as defining with greater frequency than the other properties. Thus, in Table 2, the association between *baseball* and "round" is so strong (70) that "round" would become the vivid property of both *button* and *hailstone.* As a result of this spread of effect, "round" would become the obvious property to define concept, "X." Investigations by Freedman and Mednick (1958), Underwood and Richardson (1956b), and Wicklund, Palermo, and Jenkins (1964) suggest the feasibility of this hypothesis.

A third competing explanation was that

TABLE 2

EXPLANATION FOR THE SELECTION OF PROPERTIES DEFINING CONCEPT "X"

Example	Common property	Associative strength
Button	Round Small Hard White	61 15 5 5
Baseball	Round White Hard Small	70 11 10 5
Hailstone	Hard Round White Small	49 14 9 7

Note.—Total associative strength for round = 145, hard = 64, small = 27, white = 25. "Round" is selected as the property to define concept, "X", according to each of the three hypotheses.

TABLE 3
ILLUSTRATIVE SETS OF EXAMPLES DESIGNED TO TEST THE THREE HYPOTHESES

Hypothesis	Example	Common property	Associative strength	Total associative strength	Property predicted to define concept "X"
1. Total strongest property	Baton	Round Thin	6 6	Round = 123 Thin = 23	Round
	Saucer	Round Thin	49 14		
	Button	Round Thin	68 3		
2. Single strongest property	Pot	Shiny Hard	8 5	Shiny = 43 Hard = 41	Hard
	Armor	Hard Shiny	31 18		
	Badge	Shiny Hard	17 5		
3. First property of sequence	A. Button	Round Hard	68 3	Round = 89 Hard = 85	Round
	B. Thimble	Round Hard	12 10		
	C. Stone	Hard Round	72 9		
	C. Stone	Hard Round	72 9	Round = 89 Hard = 85	Hard
	B. Thimble	Round Hard	12 10		
	A. Button	Round Hard	68 3		

would determine which properties would be the order in which the examples appeared selected as defining concept, "X." Specifically, it was hypothesized that the strongest property associated with the first example *S* saw would be selected as defining with greater frequency than the strongest property of succeeding examples. Thus, in Table 2, since *S* would most likely see *button* first and since its strongest property is "round" (associative strength = 61), the latter would be verified as a property of *baseball* and *hailstone,* and then be given as defining concept "X." (Experimentally, "round" would never be the strongest property of any of the other examples.) Studies by Cohen and Musgrove (1964), Coleman (1963), Crouse and Duncan (1963), and Freedman and Mednick (1958) indicate the reasonableness of this hypothesis.

METHOD

One hundred and seven[2] of the 213 nouns (examples) in the Underwood and Richardson norms were presented, one at a time, to 232 fifth-grade children in the Palo Alto Unified School District, to determine what properties children would as-

[2] If adult norms had been used in the present study, combinations of these 107 examples would have provided optimal testing of the three hypotheses. It was assumed, that these same examples would generally provide optimal examples when the children's norms had been established.

sociate with each example as well as their respective associative strengths. Each example was presented visually for 10–15 seconds and *S*s were asked to think of another word which best described the word in front of them. "Make sure the word you think of tells how something smells, or how it tastes, or what size it is, or something like that." Several words were presented as examples and various appropriate and inappropriate response words were discussed until *S*s fully understood their task. All properties associated with each example as well as their frequencies of occurrence were tabulated. Thus, "round," 68%; and "small," 17% (miscellaneous, 15%) were the two major properties associated with the example *button*. Since 68% of the students responded with "round," it represented the stronger of the two properties.

Examples with at least two common properties were then grouped into sets with three examples per set to test one of the three hypotheses (Table 3). To qualify for testing the total-strongest-property hypothesis, a set had to be arranged as follows: The first example had about equal associative strengths for both properties (to control for a first-property effect), and the total strength for the two properties differed maximally (Hypothesis 1, Table 3).

To qualify for testing the single-strongest-property hypothesis, a set had to be arranged as follows: The first example had nearly equal associative strengths for both properties (control for a first-property effect), and the total strengths were nearly equal (control for total-strongest-property effect). The second and third examples in the set each had properties which differed in their strengths to allow one of the examples to have a very strong property. Thus, in Hypothesis 2, Table 3, "hard" is a relatively strong property of armor (31%).

To qualify for testing the first-property hypothesis, a set had to be arranged as follows: The first example had a high strength for one of its properties and a low strength for the other; the second example had about equal strengths for both properties, and the last example had a high strength for one property and a low strength for the other (which was a reverse of the first example; see Hypothesis 3, Table 3). If the set were presented in the sequence A, B, C, then one property (round) should be selected; if however, the set was presented in the sequence C, B, A, the other property (hard) should be selected. Since the single strongest property and the total strongest property were held constant across sequences, confounding effects were partially controlled.

Of the 39 sets which qualified for testing Hypothesis 1, the 20 most powerful were selected. Of the 43 sets which qualified for testing Hypothesis 2, 16 of the strongest were selected, and of the 33 for Hypothesis 3, 17 were selected and duplicated with the exception that the first and third examples were reversed. The total number of sets of examples designed to test the three hypotheses was 53.

Each of 222 different fifth graders in the same school district was presented with one-half the sets designed to test each of the three hypotheses. Each experimental *S* received 10 for Hypothesis 1, 8 for 2, and 17 for 3, totalling 35 sets. Eight different orders resulted from all possible combinations of half the sets for each hypothesis; these orders were randomly assigned to the 222 *S*s. Each set (three examples) was printed on a separate card and presented one at a time for 10–15 seconds. Directions were similar to those given when one example was presented with the exception that, in this situation, *S*s were to think of a property describing all three examples on the card.

RESULTS

The mean number of *S*s selecting a property from each set of examples was 91.7. *Baton, saucer,* and *button* was one of the sets designed to test the total-strongest-property hypothesis. Of *S*s responding to this set, 95.7% selected the property "round" (predicted property, see Hypothesis 1, Table 3) and 4.3% selected the property "thin." These percentages were placed in a 2 × 2 table as observed percentages along with expected percentages of 50% "round" and 50% "thin," and a chi square was computed. This same procedure was followed for each set designed to test Hypotheses 1 and 2. Chi squares for Hypothesis 3 were slightly different. Frequencies rather than percentages were used, and they occurred for sequence A, B, C and for sequence C, B, A, rather than as observed and expected frequencies.

For each set testing each hypothesis, two questions were important: (*a*) Did the greatest percentage of *S*s select the predicted property? (*b*) Were the differences in percentages significant? (Hypotheses 1 and 2, Table 3).

Hypothesis 1 predicted that, in each of the 20 sets, the total strongest property would be selected by the greatest percentage of *S*s. Table 4 shows that, for 18 of the 20 sets, the greatest percentage of *S*s did select the predicted property; in 17 of the 18 sets, the difference between the percentage selecting the predicted property and the percentage selecting the opposite property was significant (for 15 of the 17, $p < .001$). Clearly, Hypothesis 1 was supported.

TABLE 4
Summary of Results

Item	Total strongest property	Single strongest property	First property
Number of sets testing hypothesis	20	16	17
Greatest percentage selecting predicted property	18	7	8
Number of chi squares			
$p < .05$	17	1	1
ns	1	6	7
Greatest percentage selecting opposite property	2	9	9
Number of chi squares			
$p < .05$		4	
ns	2	5	9

Hypothesis 2 predicted that, in each of the 16 sets, the single strongest property would be selected by the greatest percentage of *S*s. Results showed that in only 7 of the 16 was the predicted property selected by the greatest percentage of *S*s, and in 6 of these the differences were not significant. Further, in 9 sets, the opposite property was selected by the greatest percentage of *S*s. Hypothesis 2 was not supported.

Hypothesis 3 predicted that (*a*) the strongest property of the first example would be selected by the greatest percentage of students, and (*b*) when the first example was replaced by a different example whose strongest property differed from the first, the greatest percentage of *S*s would select the "new" strongest property. Results showed that this occurred in only 8 of the 17 situations; in 7 of the 8, the differences in frequencies were not significant. In the remaining 9 situations, little or no change occurred. Hypothesis 3 was not supported.

Discussion

A conclusion that can be drawn from the results is that when one of two properties is to be selected as defining some concept for which *S* is given verbal examples, the property with the highest total associative strength will be selected. Neither the order in which the examples are presented nor the presence of an example with a strongly associated property appears to influence the selection of the defining property.

Several studies (Crouse & Duncan, 1963; Freedman & Mednick, 1958; Judson & Cofer, 1956) have shown that the order in which verbal examples appear influences selection performance on verbal concept tasks. These studies used adult *S*s. The present study, using children as *S*s, failed to obtain similar results. One possible explanation for this inconsistency is that age, in part, determines when the order of example presentation will affect the selection of defining properties. Such an explanation would be more convincing if the examples and selection tasks had been similar over the previous experiments. In one study, the concept examples were names of processes; in the other two, they were names of objects. In one study, the selection task was that of sorting cards after being presented with a positive example; in another, the task was that of discarding one of four words which did not go with the other three. In the third study, the task was identifying examples of several different concepts presented in a serial order. Given this variety of selection tasks and the differing nature of the examples, conflicting results about the effects of example sequence are probably more attributable to example and task differences than to age differences.

This same explanation may also account for the conflicting results on the effects of a single strong property on the selection of defining properties. The study by Freedman and Mednick (1958) differs from the present study in the nature of the selection task and the methods of example presentation.

The present study dealt with a simple concept task. Only three examples were used from which to select defining properties. Furthermore, the study was designed so that only one property could be selected and only two were common to each example. A next logical step would be to select concepts more similar to those presented in the classroom which characterize a wider variety of examples and a greater number of common properties.

REFERENCES

COHEN, J. C., & MUSGROVE, B. S. Effect of meaningfulness on cue selection in verbal paired-associate learning. *Journal of Experimental Psychology,* 1964, **68,** 284–91.

COLEMAN, E. B. The association hierarchy as an indication of extraexperimental interference. *Journal of Verbal Learning and Verbal Behavior,* 1963, **2,** 417–21.

COLEMAN, E. B. Verbal concept learning as a function of instructions and dominance level. *Journal of Experimental Psychology,* 1964, **68,** 213–14.

CROUSE, J. H., & DUNCAN, C. P. Verbal concept sorting as a function of response dominance and sorting method. *Journal of Verbal Learning and Verbal Behavior,* 1963, **2,** 480–84.

FREEDMAN, J. L., & MEDNICK, S. A. Ease of attainment of concepts as a function of response dominance variance. *Journal of Experimental Psychology,* 1958, **55,** 463–66.

JUDSON, A. J., & COFER, C. N. Reasoning as an associative process: I. "Direction" in a simple verbal problem. *Psychological Reports,* 1956, **2,** 469–76.

SCHULZ, R. W., MILLER, R. L., & RADTKE, R. C. The role of instance contiguity and dominance in concept attainment. *Journal of Verbal Learning and Verbal Behavior,* 1962, **1,** 432–35.

UNDERWOOD, B. J., & RICHARDSON, J. Some verbal materials for the study of concept formation. *Psychological Bulletin,* 1956, **53,** 84–95. (a)

UNDERWOOD, B. J., & RICHARDSON, J. Verbal concept learning as a function of instructions and dominance level. *Journal of Experimental Psychology,* 1956, **51,** 229–38. (b)

WICKLUND, D. A., PALERMO, D. S., & JENKINS, J. J. The effects of associative strength and response hierarchy on paired-associative learning. *Journal of Verbal Learning and Verbal Behavior,* 1964, **3,** 413–20.

(Received July 17, 1967)

Journal of Educational Psychology
1968, Vol. 50, No. 5, 334–338

CHANGES IN PUPIL ATTITUDES DURING THE SCHOOL YEAR[1]

NED A. FLANDERS
University of Michigan

BETTY M. MORRISON
Western Reserve University

AND E. LELAND BRODE
Northern Illinois University

Data from 2 separate studies indicate a significant loss in positive attitudes of pupils toward their teachers and schoolwork during the school year. In the present study of 820 6th-grade pupils in 30 classrooms it was shown that this erosion of positive attitudes is not related to pupils' IQ, socioeconomic status, or percentage of A and B letter grades assigned by the teacher, but is related to the "externality" or "internality" of the pupils and to the teachers' verbal classroom behavior. Greater losses in attitudes occurred among external than among internal pupils and among pupils whose teachers exhibited a lower incidence of praise and encouragement than among those whose teachers exhibited a higher incidence of such behaviors.

Are pupils most optimistic about schoolwork as the school year begins? Does this optimism erode as the school year progresses? Are there particular patterns of teacher behavior which appear when there is less erosion of optimistic pupil attitudes? This article will attempt to answer these questions, not with complete, unequivocal answers, but with some suggestions based on two separate studies.

In a 1960–61 Minnesota study (Flanders, 1963), fairly conclusive evidence was collected indicating that over 3,000 students in two junior high schools scored highest on an attitude inventory assessing positive perceptions of their teachers and their schoolwork in October, only to have a statistically significant decrease in the scores of a January readministration of the same inventory. A follow-up administration was about the same as January, significantly lower than the October scores.

The 1960 attitude inventory consisted of 59 items roughly divided into four subscales on the basis of content: (*a*) teacher attractiveness, which included such items as, "I would like to have this same teacher next year," and, "This is the best teacher I ever had," (*b*) fairness of rewards and punishments, which included such items as, "This teacher punishes me for things I didn't do," and, "This teacher punishes the whole class when he (she) can't find out who did something," (*c*) teacher competence, which included such items as, "Our teacher is very good at explaining things clearly," and, "It is easy to fool this teacher," and (*d*) interest in schoolwork, including such items as, "This teacher makes everything seem interesting and important," and, "Most of us get pretty bored in this class." The response to each item was on a 5-point scale from strongly disagree to strongly agree. All items were keyed so that a higher score represented more positive attitudes and perceptions.

The mean score of the October administration of the attitude inventory was 217. The means of the January and May administrations were 204 and 205, respectively—both significantly lower ($p < .01$) than the October administration. These data were collected in so-called academic classes in Grades 7, 8, and 9, excluding such subjects as physical education, music, home economics, and shop.

The results seem quite clear. There is a significant reduction in the average scores of positive pupil perceptions between October and January of the school year.

The Present Study

During the 1964–65 school year a Michigan Student Questionnaire (MSQ) was

[1] This article is based on research supported by the United States Office of Education. The first author was project director; the second author was in charge of statistical analysis; and E. Leland Brode was in charge of data collection.

administered to 101 sixth-grade classes in 15 school districts near Ann Arbor. Thirty classes were selected for further study from the October distribution to include the top 10, the bottom 10, and 10 near the average of the 101 classes. The test was readministered in January and again in May in these 30 classes, each administration involving more than 800 pupils, and the sample can be considered representative of over 3,000 pupils who were in the larger population.

The MSQ was essentially the same inventory used in 1960–61 except that the items had been simplified in an effort to adjust the vocabulary to the reading skills of sixth-grade pupils. A factor analysis of the MSQ indicated that the most important factor was teacher attractiveness, with additional factors of teacher competence, teacher fairness, and lack of pupil anxiety forming a combination which was less important than the first factor.

The means for each of the 30 classes are shown in Table 1, arranged in terms of the top, middle, and bottom 10 classes on the October administration.

The 1964–65 results were nearly identical to the 1960–61 results. There was a significant drop in average scores of positive pupil attitudes during the first 4 months of the school year. The mean score for the October administration was 178.2 with a standard deviation of 26.52. The January administration had a mean of 172.2 and a standard deviation of 31.13; and the May administration had a mean of 170.6 and a standard deviation of 30.60.

The rest of this article will discuss the various factors that might be related to the observed change in attitude, one administration compared to the next.

Factors Not Associated with Change in Attitude

Simple regression is not an adequate explanation of these changes. While it is true that three low classes (Classes 21, 24, and 27) showed the highest positive changes, it is also true that three other low classes (Classes 22, 23, and 28) showed large decreases. The average loss (October to May) of the bottom 10 classes was 6.2, compared to 7.6 for the total group. Furthermore, the top 10 classes do not show uniform loss but instead are symmetrically distributed about an average loss of 7.5.

The correlations between administrations, based on individual scores, were positive and fairly high: for October to January, $r = .704$; for January to May, $r = .812$; and October to May, $r = .655$. The correlation between October and May, based on 30 class averages, is higher; $r = .876$. The correlations present a picture of a fairly stable response pattern both within and between classes.

There is the possibility that change in positive pupil attitudes might be associated with the average class IQ, socioeconomic status, or the percentage of A and B letter grades assigned to the pupils by the teacher. Table 2 shows such data for the nine classes which had high change losses and the seven classes with the least

TABLE 1

MEANS OF 30 CLASSES ON THE 1964–65 MICHIGAN STUDENT QUESTIONNAIRE

Class	Administration			Class	Administration			Class	Administration		
	Oct.	Jan.	May		Oct.	Jan.	May		Oct.	Jan.	May
1	204.9	194.7	194.2	11	180.9	163.7	156.3	21	166.4	173.1	175.0
2	201.4	200.0	195.4	12	179.2	180.3	176.2	22	166.2	147.7	150.4
3	200.3	204.9	200.0	13	178.8	173.3	164.5	23	165.1	151.3	143.8
4	199.6	194.7	192.3	14	176.8	171.4	155.7	24	162.7	174.7	173.9
5	197.5	195.2	195.2	15	175.2	151.9	155.7	25	162.7	159.8	160.8
6	197.0	193.5	189.9	16	178.8	178.0	176.1	26	158.1	149.4	147.2
7	193.6	190.6	187.9	17	177.9	185.4	178.2	27	158.1	154.0	161.2
8	192.0	185.7	175.9	18	176.2	166.6	169.3	28	157.2	145.5	139.2
9	191.6	178.9	186.5	19	175.4	169.2	167.0	29	156.6	137.3	147.7
10	190.3	178.3	176.3	20	173.8	170.2	166.5	30	149.9	137.2	142.2

TABLE 2
Comparisons between High- and Low-Attitude Change Classes

High change				Low change			
Class no.	IQ	SES rating	% of A & B grades	Class no.	IQ	SES rating	% of A & B grades
11	121.6	81	75	3	118.1	65	91
23	104.2	67	33	17	114.1	78	55
14	111.7	71	64	25	104.1	69	48
15	118.5	67	52	5	113.1	70	66
28	109.8	74	41	16	120.7	71	81
8	116.8	78	57	12	100.0	63	50
22	110.3	67	63	27	116.1	83	62
13	116.0	83	80				
10	112.7	70	42				

Note.—SES = socioeconomic status.

amount of change. The mean IQ for the high-change group was 113.5, while it was 112.3 for the low-change group. Here the IQ scores used were those based on school records and probably involved different published tests. The median socioeconomic rating for the high-change group was 71; a median rating of 70 was obtained for the low-change group. Here a rating on the National Opinion Research Center scale (Reiss, 1961) was made of the wage earner's occupation as reported by the teacher. The mean percentage of A and B letter grades for the high-change group was 56.5 and for the low-change group, 64.5. While this last difference is consistent with a theory that change involving loss of positive attitudes is associated with receiving lower grades, a z test between independent proportions yielded a value of 1.66 which was not high enough to reject the null hypothesis at the .05 level of significance. All of these data suggest that changes in class attitudes are not significantly associated with average IQ, socioeconomic status, or grades given by the teacher.

Two Factors Associated with Change in Attitude

In another study, Morrison (1966) has shown that Rotter's notion of "externality" and "internality" (Rotter, Seeman, & Liversant, 1962) can be assessed among sixth-grade pupils. By externality is meant the tendency of a pupil to believe that his successes and failures are caused by forces beyond his control. By internality is meant the tendency to believe that successes and failures are self-determined and products of one's own behavior. External children, according to Morrison's conception, would be more likely to associate the good and bad outcomes of classroom learning activities with the teacher who is a powerful source of influence. Internal children, on the other hand, would see themselves as more closely associated with the good and bad characteristics of learning outcomes.

A test of internality-externality was administered to all the pupils in the 30 classes during the January administration of tests. The test consisted of 26 items, each containing two statements, and the pupils responded by marking the statement in each item which they believed was more often true. Typical items were: (*a*) "If you study you will do well on a test," or (*b*) "People who score the highest on a test are lucky."—(*a*) "Most of the time children get the respect they deserve from others," or (*b*) "Many times a child can try hard and no one will pay attention to him."—(*a*) "Usually other people choose me for a friend," or (*b*) "Usually I choose my own friends."—(*a*) "Children get into trouble because their parents punish them too much," or (*b*) "The trouble with most children is that their parents are too easy with them."

Each item was scored 1 if the internal response was chosen and 2 if the external response was selected, giving a possible range from 26 to 52 for the total scores. The actual scores ranged from 26 to 49. Students in the lower third of the distribution (raw scores of 31 and below) were defined as internals and those in the upper third (raw scores of 36 and above) were defined as externals.

In addition to these tests each of the 30 classrooms was visited by an observer trained to code verbal communication into the 10 categories of interaction analysis developed by Flanders (1965). More than six visits were made to each class and more than 7,000 tallies were recorded by observers. The main results of interaction

analysis will be reported elsewhere; the interest for the moment is in the incidence of praise and encouragement expressed by the teachers during these visits. The occurrence of this type of teacher statement varied from low of .2% to a high of 2.1% of all tallies recorded by the observer. The problems of reliability among observers and the representativeness of the interaction sampled are too complex to be discussed here. It can be said, however, that the relative objectivity of the observation data, or lack of it, would affect the data from all classes equally and cannot account for any of the differences about to be discussed.

It was hypothesized in this study that:

1. External children have a greater negative shift in attitude than do internal children.

2. The classes of low-praise teachers have a greater negative shift in attitude than do the classes of high-praise teachers.

3. The attitudes of external children are more affected by the praise and encouragement of the teacher than are the attitudes of internal children.

To test these hypotheses a two-way analysis of covariance in the case of unequal or disproportionate numbers of observations in the subclasses was performed using the third attitude inventory scores as the dependent variable and adjusting with the scores from the first administration. Table 3 includes the October means, the May means, and May means adjusted for the initial attitudes, and the change means. These are arranged in subgroups of internal and external pupils and pupils with high-praise and low-praise teachers.

The slope of the regression line was .845, and the analysis of covariance yielded an error mean square of 520.7. The main effect for internal versus external pupils produced a mean square of 18,331 and the resulting F ratio is significant at well beyond the .01 level ($F = 35.20$, $df = 1/473$). The main effect for pupils of high-praise versus low-praise teachers produced a mean square of 7,128, also resulting in an F ratio which is significant at well beyond the .01 level ($F = 13.69$, $df = 1/473$).

TABLE 3

INITIAL, FINAL, AND ADJUSTED MEANS OF ATTITUDE SCORES FROM ANALYSIS OF COVARIANCE BY PUPIL TYPE AND TEACHER STYLE

Teacher style	Pupils		
	Internal	External	All
High-praise			
Initial M	190.6	171.0	183.9
Final M	187.3	159.2	177.7
Adjusted M	163.7	152.1	159.7
Change M	−3.3	−11.8	−6.2
Low-praise			
Initial M	178.7	121.5	143.9
Final M	173.1	111.3	135.5
Adjusted M	159.5	146.0	151.3
Change M	−5.6	−10.2	−8.4
All			
Initial M	185.8	137.8	162.6
Final M	181.6	127.1	155.2
Adjusted M	162.0	148.0	—
Change M	−4.2	−10.7	—

However, the interaction of pupil types and teacher styles resulted in a mean square of only 90, which is not significant ($F = .17$).

These results indicate that not only did external pupils have less positive attitudes than did internal pupils early in the school year, but when the May scores are adjusted by the October scores, it is apparent that external pupils experienced significantly greater declines in their attitudes than did internal pupils. Also, pupils with low-praise teachers showed greater losses in positive attitudes during the year than did pupils with high-praise teachers. However, there was no evidence that the attitudes of external children were more affected by praise or lack of praise on the part of the teacher than were the attitudes of the internal children.

DISCUSSION AND CONCLUSIONS

In two separate projects the attitude inventory scores indicate that positive perceptions of pupils toward their teacher and their class activities decrease sometime during the first 4 months of the school year. In the second project these changes were shown to be unrelated to IQ, performance grades assigned by the teacher, and the socioeconomic ratings of the

father's occupation. Two hypotheses about changes in attitude were supported. First, external pupils experience a greater loss of positive attitudes toward school than do internal pupils. Second, in classrooms of teachers who provide less praise and encouragement there is greater loss of positive attitudes than in classrooms of teachers who provide more praise and encouragement. The third hypothesis was not supported since the interaction effects in Table 3 are not significant.

One inference to be drawn is that the type of youngster who is more dependent on external influences seems to be more likely to suffer a loss of positive expectations than is the one who is more dependent on internal influences. In addition, pupil attitudes toward the teacher and the learning activities seem to be related to teacher behavior. Whether this difference in pupil attitudes is the result of the different teacher behaviors, or the different amounts of teacher praise and encouragement are the result of the pupils being more or less deserving of that praise, is not clear from the evidence of the present study. The absence of a significant difference between high-change and low-change classes with respect to the percentage of A and B grades given by the teacher (Table 2) would indicate that the pupils' performance was not the deciding factor. Also, previous studies (Flanders, 1963, 1965) have indicated that teacher behavior is the more dominant factor and that differences in such patterns of teacher influence tend to be greater between different teachers than between different situations for the same teacher.

In this sample, differences among the pupils had a greater effect than the presence or absence of a small amount of teacher praise and encouragement. Future studies of the erosion of positive pupil attitudes may wish to take into account other differences among pupils as well as differences in teacher behavior.

This study did not provide direct evidence concerning two opposite hypotheses about the erosion of positive pupil attitudes. One theory is that the pupils become disenchanted with the teacher during the first few months of the school year. A second theory, based on the assumption that the October scores are inflated or too high, is that as the pupils learn to trust their teacher they do not overestimate their ratings as they felt compelled to do with a strange teacher. Without going into detail, the authors tend toward the first of these two theories, primarily because the teacher's behavior is the predominant influence in the typical classroom; but much more evidence will be required before any conclusions can be reached.

Meantime, lack of loss of positive attitudes may be the mark of a good match between teaching behavior and particular attributes among pupils. Apparently, in most classrooms such a match does not exist.

REFERENCES

Flanders, N. A. Helping teachers change their behavior. Terminal report, National Educational Defense Act, Title VII project, 1963.

Flanders, N. A. *Teacher influence, pupil attitudes, and achievement.* United States Office of Education Cooperative Research Monograph No. 12. Washington, D. C.: United States Government Printing Office, 1965.

Morrison, B. M. The reactions of external and internal pupils to patterns of teacher behavior. Unpublished doctoral dissertation, University of Michigan, 1966.

Reiss, A. J., Jr. *Occupations and social status.* New York: Free Press of Glencoe, 1961.

Rotter, J. B., Seeman, M., & Liversant, S. Internal versus external control of reinforcements: A major variable in behavior theory. In N. F. Washburne (Ed.), *Decisions, values, and groups.* Vol. 2. London: Pergamon Press, 1962.

(Received July 21, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 5, 339–349

DIRECT VERSUS VICARIOUS PAIRED-ASSOCIATE LEARNING IN RETARDATES AND NORMALS[1]

SEYMORE SIMON AND RAYMOND DITRICHS[2]
Northern Illinois University

In 4 experiments retardate and normal *Ss* either listened to a peer learning an initial list of PAs or learned the same list directly prior to receiving a PA test list. The major variables studied were transfer paradigm (A-B, A-B; A-B, A-C; A-B, A-B′; A-B, C-D), number of 1st-list trials (5, 9, 15), and instructions to observers ("listen" or "learn"). Results indicated that for both populations: (*a*) vicarious List-1 learning facilitated transfer in the A-B, A-B paradigm, the extent of this transfer depending upon instructional conditions for normals but not for the mentally retarded, and (*b*) there was a consistent trend for performance to be better when List 1 was learned directly rather than vicariously. The results were taken as support for the notion that retardates are more outer-directed than normals and the failures to obtain significant transfer following vicarious learning in the A-C and A-B′ paradigms were attributed to low levels of initial learning. It was suggested that vicarious verbal learning is best conceptualized within the framework of traditional PA transfer theory.

It is now well established that a wide range of behaviors can be learned in human adults and children by vicarious means. These include galvanic-skin-response aversive conditioning (Berger, 1962), verbal response classes (Ditrichs, Simon, & Greene, 1967; Kanfer & Marston, 1963; Marston, 1964, 1966; Simon, Ditrichs, & Jamison, 1965), nonverbal aggression (Bandura & Walters, 1963), and discriminative responding (McDavid, 1959; Paschke, Simon, & Bell, 1967).

A common assumption in such research is that the same fundamental learning parameters are operative in both direct and vicarious acquisition, for example, frequency, contiguity, etc. However, the extent to which the social learning situation introduces additional variables which may facilitate or hinder performance relative to direct procedures remains a continuing theoretical and empirical question. That the former can occur, for example, has been demonstrated by Bandura and McDonald (1963) who found that children observing adult models expressing moral judgments counter to the group's orientation showed significant changes in their judgments relative to *Ss* who were directly reinforced for the same deviant behavior. Similarly, Elkonin (1957) reports the results of a study by Pen who investigated the vicarious acquisition of positive and inhibitory conditioned reflexes in children. Quoting Pen, it is noted that "in many cases where a conditioned connection could not be produced in the child in the usual way (by timing a signal just before, or at the same time as, reinforcement) it sprang up easily by imitation [p. 62]." By contrast, other research has demonstrated slower acquisition under observational than under direct conditions (e.g., Van Wagenen & Travers, 1963).

An analysis of the factors responsible for either the facilitation or inhibition of learning under observational conditions may be of some importance in suggesting possible reasons for observed performance differences between populations of "slow learners," for example, retardates, and so-

[1] This research was supported by funds from United States Public Health Service Grant MH-08917. The authors would like to express their appreciation to the DeKalb, Elgin, and Rochelle public school systems, and to the Psychology Department of Dixon State School, for their help in providing *Ss* and facilities. The authors are especially indebted to Rolland Metzger of the Dixon State School for his invaluable assistance. Preliminary findings of Experiments I and II were presented at the Midwestern Psychological Association Meetings, Chicago, 1964.

[2] Thanks are due Richard E. Paschke, Charles C. Holmes, John Klinefelter, and Cherry Alborg, who assisted in various stages of this research.

called "normal" populations. For example, it has been hypothesized that attentional deficits are primarily responsible for the inferior performance of the mentally retarded on selected tasks (Zeaman & House, 1963). Since it is explicitly assumed that attentiveness to the model's behavior is a necessary component of observational learning, and since retardates have been characterized as being more outer-directed than normals (Zigler, 1966), one way in which the observational situation may enhance learning is by facilitating attentional responses to the task-relevant cues. That essentially the same interpretation is given in educational research to account for the learning superiority of overt as against covert rehearsal of materials (Anderson, 1967), however, suggests that both direct and vicarious procedures may facilitate attentional responding. More important, these conflicting interpretations indicate that the conditions under which one is a more effective learning technique than the other have not yet been delineated.

Among other deficits, an impoverished verbal repertoire has frequently been said to characterize the retardate individual. This is implied in the difficulty retardates have in coordinating motor and verbal behavior, semantic generalization anomalies, and lowered verbal learning acquisition rates (Denny, 1964). In line with previous discussion, then, a potentially fruitful area of research would be the systematic study of direct versus vicarious verbal acquisition in retardate populations. The present experiments were specifically concerned with this problem and were designed to compare the direct with the vicarious learning of paired-associate (PA) verbal materials. Both normals and retardates were employed as *S*s, and treatment groups were constituted to examine direct transfer (stimuli and responses identical in training and test lists), associative facilitation and interference, and learning-to-learn effects, under direct and vicarious conditions of first-list learning.

EXPERIMENTS I AND II

Method

Subjects. In Experiment I, 75 normal children were selected from three sixth-grade classes of a local junior high school. The three classes represented low, average, and high reading ability children, respectively, as determined by reading comprehension scores on the Iowa Test of Basic Skills. In Experiment II, 30 institutionalized male familial retardates served as *S*s. In both Experiments II and IV, *S*s with known organic brain damage and histories of epilepsy were excluded from the sample. Information concerning race and socioeconomic status was not obtained. Descriptive data for all *S*s are presented in Table 1.

Procedure. Several weeks prior to the treatments of Experiment I, children in each of the three classes were asked to vote for a member of their class to whom they would listen most if he or she were allowed to be the teacher for a day. The child receiving the greatest number of votes in each of the classes, two girls and a boy, then made four tape recordings with *E*, a male graduate student. In each of these recordings the selected child acted as an *S* in a simulated PA learning task. The materials presented to the taped *S* were six bigram-word PAs. Stimulus bigrams were of low intralist similarity and were selected from norms provided by Underwood and Schulz (1960). Response words

TABLE 1

MEANS, STANDARD DEVIATIONS, AND RANGES OF AGES, INTELLIGENCE, INSTITUTIONALIZATION, AND READING COMPREHENSION SCORES

Experiment	Subjects	*N*	Age (yrs.)			Intelligence[a]			Reading comprehension[b]			Institutionalization (yrs.)		
			M	*SD*	Range	*M*	*SD*	Range	*M*	*SD*	Range	*M*	*SD*	Range
I	Normals by reading level													
	Low	25	11.8	0.6	11.0–13.5	98.3	13.8	77–128	4.8	0.9	3.0–6.7			
	Average	25	11.3	0.4	10.9–11.8	106.4	10.6	87–126	6.8	1.0	5.2–8.6			
	High	25	11.4	0.3	10.9–12.3	120.1	8.4	103–131	8.7	1.3	6.7–10.7			
II	Retardates	30	15.4	1.6	13.0–18.0	64.4	8.0	50–84				7.2	3.4	2.2–13.1
III	Normals	144	10.2	1.1	9.1–11.3	112.2	10.1	101–129						
IV	Retardates	96	15.0	0.9	11.2–18.1	52.2	9.4	42–67				5.8	2.1	2.0–14.2

[a] Based on Otis, WISC, or Kuhlmann-Anderson tests for normals and on Stanford-Binet, 1937 Rev., WISC, WAIS, and Merrill Palmer tests for retardates. Since these scores were taken from school or institutional records and were based upon numerous tests, mental ages are not reported.

[b] Grade-equivalent scores.

were culled from children's word association norms (Palermo & Jenkins, 1964) and were all AA words equated for Thorndike-Lorge frequency of occurrence on the G count. Following a single presentation of the six PAs, the taped *S* responded with the correct associate following each of the six bigrams for nine successive errorless trials. Five random orders of the PAs, replicated twice, were presented to the taped *S*. The taped-presentation sequence was as follows: *E* spelled the bigrams, *S* responded with the appropriate word within a 3-second anticipation interval, *E* repeated the bigram and pronounced the correct response word. The interitem interval was 3 seconds. A 6-second intertrial interval was employed with a bell being sounded midway through this interval to signal the start of a new trial.

All *S*s were run individually in a room provided by the school. Upon entering the room, *S*s were first required to learn a three-item PA list consisting of single number–single letter pairs to a criterion of 2 successive errorless trials or for a maximum of 25 trials. All *S*s tested were able to learn these PAs within the 25-trial limit. Instructions as well as stimulus and response items were presented through headphones via tape. Following this warm-up task, each *S* was told that he was going to hear a tape recording of one of his classmates learning some letters and words. Instructions to *S* were simply to listen to the taped learning session. The *S* then heard his classmate receive PA instructions followed by one of the four prerecorded tasks. Immediately following this listening session, *S* was instructed to learn some material himself in the same way that he had just heard his classmate learn. A PA transfer task followed these instructions and *S* was required to learn this list to a criterion of 3 successive errorless trials or for a maximum of 40 trials. The transfer task was administered via tape at presentation rates identical to those employed in the vicarious session.

The experimental design called for the use of five treatment groups. For Group E-1, a direct transfer group, the stimulus and response members of the training tape were identical to those employed on the transfer tape. For Group E-2, a mediated facilitation group, the stimulus members in the training and transfer tapes were identical, but the response members of the transfer tapes were dominant word associates of the training tape responses as determined from the children's word association norms employed. For Group E-3, an associative interference group, the stimulus members of the two tapes were the same but the responses were words having no associative relationship to one another. For Group C-1, the stimulus and response members of the training and transfer tapes differed from each other. This group provided a reference level with respect to which the differential transfer effects of the experimental groups could be evaluated. A second control group, C-2, received a digit cancellation task in lieu of vicarious PA experience for the duration of the training phases of the other groups.

The PA materials to which Group E-1 listened during the vicarious phase, as well as those presented to all groups in the transfer list were: XP-*low*, MB-*chair*, EG-*fast*, KI-*sleep*, FT-*white*, and SC-*flower*. These same bigrams were paired with the following response words during the vicarious phase for Groups E-2 and E-3, respectively: *high, table, slow, bed, black, blossom; music, cheese, king, long, girl,* and *memory*. Group C-1 was exposed to the following PAs during the vicarious phase: AF-*music*, GL-*cheese*, LU-*king*, PH-*long*, RV-*girl*, and WN-*memory*.

In Experiment II, institutionalized male retardates were treated in essentially the same manner as were the normal *S*s with the following exceptions: (*a*) The *S*s were asked to vote for a member of their cottage to whom they would listen most if he were allowed to be a teacher for a day. The male retardate selected then served as *S* in all tape recordings of the major experimental conditions; (*b*) the experiment was conducted in a mobile trailer situated on the institution grounds; (*c*) a female *E* was used; (*d*) a 2,000 cps. tone was sounded 1 second prior to the onset of each stimulus bigram throughout the training and transfer tasks. A pilot study had revealed that such a procedure was necessary to hold the attention of the retardates during the aurally presented PAs. Other aspects of the procedure, including the stimulus materials employed, rate of presentation, etc., were identical to those employed with normal *S*s.

Results

The major dependent variables were mean correct responses and trials to criterion. Since essentially comparable findings were obtained with the two measures (for normals, $r = -.85$; for retardates, $r = -.90$), only results employing the former measure will be presented.

Experiment I. Figure 1 shows the mean number of correct anticipations on the criterion list by blocks of five trials for all groups. A repeated measurements analysis of variance on Blocks 1 and 2 was first conducted to obtain a measure of group differences unconfounded by possible ceiling effects, that is, *S*s in all groups reaching criterion. The results of this analysis indicated that Reading Levels ($F = 6.18$, $df = 4/60$, $p < .01$) and Blocks ($F = 151.59$, $df = 1/60$, $p < .001$) were significant sources of variance. The Groups effect fell short of statistical significance ($F = 2.04$, $df = 4/60$, $p < .10$). None of the interactions was statistically reliable. As can be seen in Figure 1, there is a tendency for *S*s in Group E-1 to give more correct responses than *S*s in Group C-2, with all

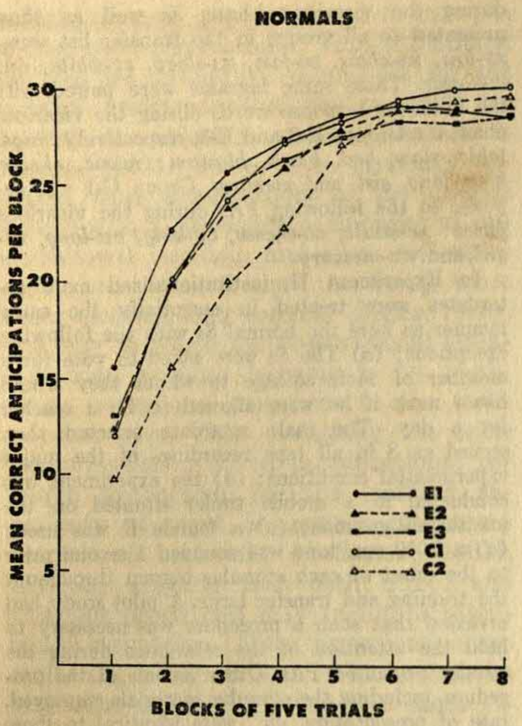


FIG. 1. Mean number of correct anticipations per block of five trials on the paired-associate test list for normal *S*s (Experiment I).

other treatment group comparisons yielding correspondingly higher p values. Tukey's test of multiple comparisons indicated that *S*s in the high reading level groups ($M = 3.90$) gave more correct responses than *S*s in the average ($M = 2.97$) and low ($M = 2.79$) reading level groups ($p < .05$ and $p < .01$, respectively); the latter two groups did not differ significantly from each other. An analysis over Blocks 1–4 yielded results essentially the same as those obtained over Blocks 1 and 2.

Error analyses over Blocks 1 and 2 were performed separately for omissions and intralist intrusions. For the omissions analysis, Reading Levels and Blocks were the only significant sources of variance ($F = 3.18$, $df = 2/60$, $p < .05$; $F = 180.55$, $df = 1/60$, $p < .001$, respectively). Tukey tests indicated that *S*s in the high reading level group ($M = 3.72$) made significantly fewer omissions than *S*s in the middle ($M = 6.28$) and low ($M = 6.54$) reading level groups ($p < .05$), with other differences failing to reach statistical significance. There were reliably fewer omissions made on Block 2 ($M = 3.20$) than on Block 1 ($M = 7.90$) of the test list. A similar analysis on intralist errors indicated that Blocks was the only significant source of variance ($F = 24.05$, $df = 1/60$, $p < .001$). The means for Blocks 1 and 2 were 8.70 and 6.59, respectively.

To ascertain the relative transfer effects of vicarious versus direct PA experience, the performance of Group E-1 on Trials 1–5 of the transfer list was compared with the performance of Group C-2 on Trials 10–14. Since Group C-2 received direct learning experience on the criterion list for Trials 1–9 and Group E-1 received vicarious experience for nine trials prior to the transfer phase, this comparison appropriately tests for the relative transfer effects of vicarious versus direct experience early in criterion list learning. A factorial analysis of variance (Groups × Reading Levels) on mean correct anticipations on Trials 1–5 for Group E-1 and on Trials 10–14 for Group C-2 revealed that reading levels was the only significant source of variance. While the group means indicated a tendency for the C-2 group ($M = 19.17$) to perform better than the E-1 group ($M = 15.89$), the F ratios for Groups ($F = 2.30$, $df = 1/24$, $p > .05$) and for the Groups × Reading Levels interaction ($F = .21$, $df = 2/24$) were not significant.

Experiment II. Figure 2 presents the mean number of correct anticipations on the criterion list by blocks of five trials. A repeated measurements analysis of variance on Blocks 1 and 2 indicated that only the main effect for Blocks was significant ($F = 43.14$, $df = 1/25$, $p < .001$). As with the results for normal *S*s, the Groups effect fell just short of statistical significance ($F = 2.52$, $df = 4/25$, $p < .10$), Group E-1 tending to give more correct responses than *S*s in the other four treatment groups.

Since relatively few numbers of retardate *S*s reached criterion within the 40-trial limit, a Groups × Blocks analysis of variance was also conducted on the number of correct responses over Blocks 1–8.

The results of this analysis indicated that Blocks was the only significant source of variance ($F = 75.55$, $df = 7/175$, $p < .001$).

An analysis of variance on number of omissions made over Blocks 1 and 2 indicated that only Blocks was significant ($F = 31.80$, $df = 1/25$, $p < .001$). Respectively, the means of Blocks 1 and 2 were 19.63 and 12.50. A similar analysis on intralist intrusions showed that Blocks was again the only significant ratio ($F = 5.85$, $df = 1/25$, $p < .05$). For retardate *S*s, there was a significant increase in intralist errors from Block 1 ($M = 5.13$) to Block 2 ($M = 6.97$) of the test list.

A *t* test was performed between the mean correct anticipations made on Trials 1–5 of the transfer list for Group E-1 and the corresponding performance of Group C-2 on Trials 10–14. Consistent with the results obtained for normal *S*s, there was a tendency for the C-2 group ($M = 9.17$) to make more correct responses than the E-1 group ($M = 6.67$), but these means did not differ significantly from one another ($t = 1.32$, $df = 10$, $p > .05$).

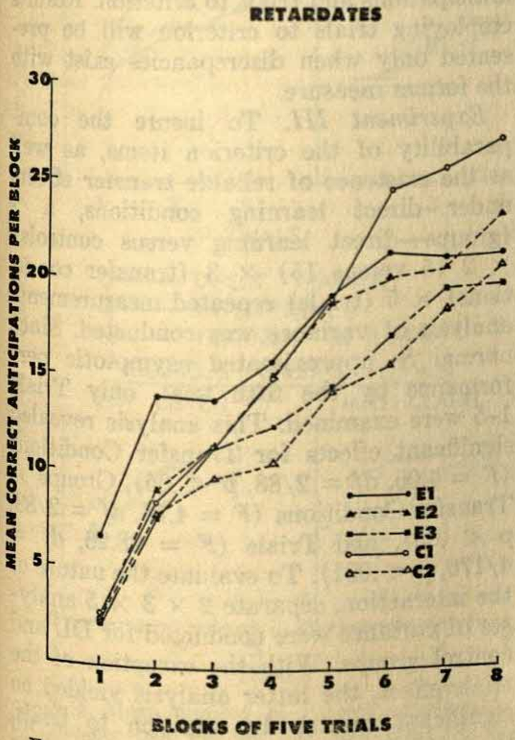


FIG. 2. Mean number of correct anticipations per block of five trials on the paired-associate test list for retarded *S*s (Experiment II).

EXPERIMENTS III AND IV

The results of Experiments I and II are equivocal in demonstrating the relative superiority of either direct or vicarious acquisition as well as in demonstrating whether vicarious PA learning occurs at all. At best, it could be said that, for both populations, direct learning produces somewhat better transfer performance than vicarious learning and that vicarious experience tends to improve performance on transfer pairs comprising the same stimuli and responses as in first-list "learning." However, both of these statements lack statistical confirmation. Consequently, Experiments III and IV were conducted to examine additional variables which may provide more definitive answers to these questions.

As stated earlier, a fundamental premise of the present research has been that, given the same task, those variables found to be significant in direct learning should also be operative under vicarious conditions. Thus, if it is assumed that the failure to find significant transfer in Experiments I and II under vicarious conditions is attributable to insufficient amounts of List-1 learning, then two variables which are likely to facilitate such learning are an increased number of observational trials and directed instructions to *S*s. That the former is a significant variable under observational conditions has been demonstrated in at least two discrimination learning studies employing retardates as *S*s (Fletcher, 1966; Paschke, Simon, & Bell, 1967). Likewise, instructions have consistently been found to be a potent learning variable in both direct (e.g., McLaughlin, 1965) and vicarious (e.g., Marston, 1966) tasks.

In Experiments III and IV, *S*s received either 5 or 15 observational trials, half under instructions to listen to the model and half under instructions to learn the materials the model was learning. The scholastic competence of the model was also manipulated as a third independent variable for normal *S*s. It was expected

that 15 trials and instructions to learn would, either as main effects or interacting variables, result in significant transfer under conditions of vicarious List-1 acquisition by increasing the degree of learning relative to Experiments I and II. To permit more precise comparisons with performance under nonobservational conditions, direct learning groups, as well as controls, were included in the design.

Method

Subjects. 154 normal children drawn from the fourth, fifth, and sixth grades of local schools and 114 institutionalized male and female familial retardates initially served as *S*s. Descriptive data for both populations is presented in Table 1.

Procedure. The general procedures in Experiments III and IV were identical and were the same as those employed in Experiments I and II with the following exceptions: (*a*) The same adult male and sixth-grade female child served as *E* and model, respectively, for all vicarious learning (VL) *S*s; this *E* also prerecorded List-1 materials for direct learning (DL) *S*s, as well as List-2 materials for all *S*s. (*b*) A mixed transfer design was employed. Stimuli for the two lists of PAs were two-digit even numbers ranging from 10 to 20. List-1 responses were *fast, sleep, soldier, eagle, king,* and *music*. Test-list interference words were *white* and *girl;* control words were *high* and *table*. (*c*) A 2,000-cps tone preceded each of the PA stimuli for all *S*s; (*d*) An 8-second intertrial interval was utilized. (*e*) The test list was learned to a criterion of 2 consecutive errorless trials or for a maximum of 35 trials. Further, to avoid score estimations for fast learners, all *S*s were run for a minimum of 10 trials on the test list.

For various reasons, for example, failure to learn the practice list, failure to cooperate at some point in the experiment, *E* mistake, 10 normal *S*s and 18 retardates were eliminated from the experiment, leaving total populations of 144 normals and 96 retardates. Those eliminated were approximately equally distributed among the various treatment conditions.

Three major independent groups were employed in each of the two experiments: VL *S*s, DL *S*s, and controls. For VL *S*s in Experiment III, the design was a 2 (instructions—"listen" versus "learn") × 2 (number of vicarious trials—5 versus 15) × 2 (model's scholastic competence—high versus low) × 3 (conditions of transfer—direct transfer, interference, learning-to-learn) mixed factorial with the first three of these variables constituting between-*S*s effects and the latter being a within-*S*s variable. The scholastic competence of the model was varied by instructing half of the *S*s that the girl learning on tape was "a very smart student and always gets very high grades." The other half of the *S*s were told that the girl was "not a very smart student and always gets very low grades." The *S*s in the "intentional" group were instructed to try to learn the words along with the taped *S* because afterwards they would be asked to do the same thing. "Incidental" *S*s were instructed simply to listen to how the girl on tape did. Twelve *S*s were randomly assigned to each of the 8 cells with the restriction that approximately equal numbers of males and females appear in each cell.

DL *S*s were required to learn List 1 directly in lieu of listening to the model's performance. To permit comparisons with VL *S*s, half of these *S*s received 5 List-1 anticipation trials while the other half of the *S*s received 15 List-1 trials prior to the test list. Finally, two groups of control *S*s played an "etch-a-sketch" game for lengths of time corresponding to List-1 learning times for DL-5 and DL-15 *S*s, respectively. There were 12 *S*s in the two DL and two control groups.

Experiment IV was identical in design to Experiment III with the exception that retardates were employed as *S*s and the scholastic competence of the model was not manipulated as an independent variable for VL *S*s. Instructions to these *S*s simply made reference to "a girl learning some numbers and words."

Results

As in Experiments I and II, the major dependent variables were mean correct anticipations and trials to criterion. Results employing trials to criterion will be presented only when discrepancies exist with the former measure.

Experiment III. To insure the comparability of the criterion items, as well as the existence of reliable transfer effects under direct learning conditions, a 2 (groups—direct learning versus controls) × 2 (5 versus 15) × 3 (transfer conditions) × 5 (trials) repeated measurements analysis of variance was conducted. Since normal *S*s approximated asymptotic performance by the fifth trial, only Trials 1–5 were examined. This analysis revealed significant effects for Transfer Conditions ($F = 4.06$, $df = 2/88$, $p < .05$), Groups × Transfer Conditions ($F = 4.78$, $df = 2/88$, $p < .05$), and Trials ($F = 38.26$, $df = 4/176$, $p < .001$). To evaluate the nature of the interaction, separate 2 × 3 × 5 analyses of variance were conducted for DL and control groups. With the exception of the trials effect, the latter analysis yielded no significant results. In addition to trials, the main effects for Transfer Conditions was significant for DL *S*s ($p < .01$), the means being 1.70, 1.48, and 1.42 for direct trans-

fer (DT), interference (INT), and new (LTL) items, respectively. Results are presented in the upper panels of Figure 3.

In the foregoing analysis, the Groups × Transfer Conditions × Number of Vicarious Trials interaction was marginally significant ($p < .10$), there being a tendency for greater DT and more INT to occur with 15 List-1 trials than with 5 trials. This trend was found to be significant in the trials to criterion analysis, the Groups × Number of Vicarious Trials × Transfer Conditions interaction being statistically significant ($F = 3.29$, $df = 2/88$, $p < .05$). As in the correct anticipations analysis, control group analysis revealed no significant ratios. A similar analysis on the DL groups, however, indicated that Transfer Conditions ($p < .05$) and the Transfer Conditions × Number of Vicarious Trials interaction ($F = 3.65$, $df = 2/22$, $p < .05$) were significant sources of variance. This interaction, along with the combined performance of control Ss, is presented in Figure 4. As can be seen, the interaction reflects the fact that significant INT effects were obtained with 15 trials on List 1 but not with 5 List-1 trials.

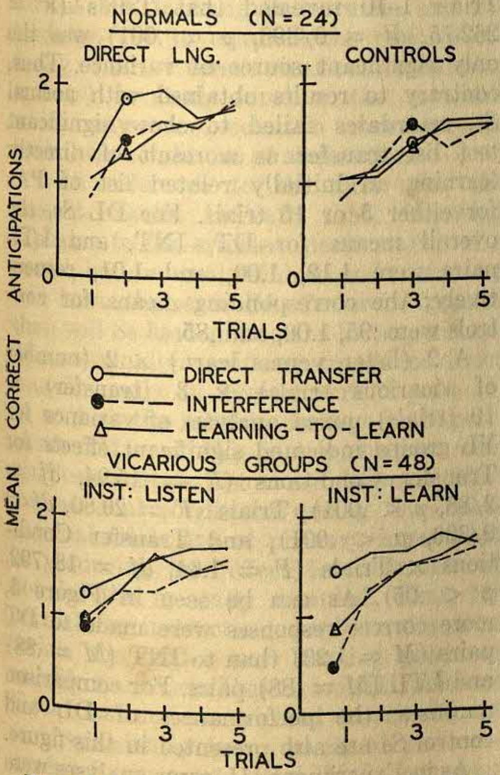


FIG. 3. Mean number of correct anticipations per transfer pair on the first five trials of the paired-associate test list for normal direct learning, vicarious learning, and control *S*s (Experiment III).

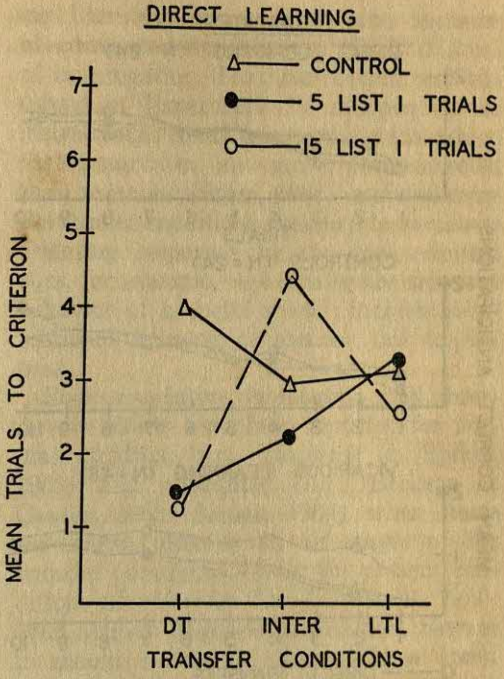


FIG. 4. Mean trials to criterion on the paired-associate test list for normal *S*s as a function of transfer conditions and number of direct List-1 learning trials (Experiment III).

Error analyses on total omissions and total intralist intrusions were essentially consonant with the correct anticipations data. For omissions, the Groups × Transfer Conditions interaction was the only significant ratio ($F = 9.23$, $df = 2/88$, $p < .001$). For DL *S*s, the means for DT, INT, and LTL pairs were 1.13, 1.96, and 1.96, respectively; for control *S*s, the means were similar and averaged 1.86. DL *S*s ($M = .60$) also made fewer intrusions than did controls ($M = 1.17$) ($F = 4.88$, $df = 1/44$, $p < .05$).

The performances of VL *S*s were next examined by conducting a 2 (listen versus learn) × 2 (number of vicarious trials) × 2 (peer status) × 3 (transfer conditions) × 5 (trials) mixed analysis of

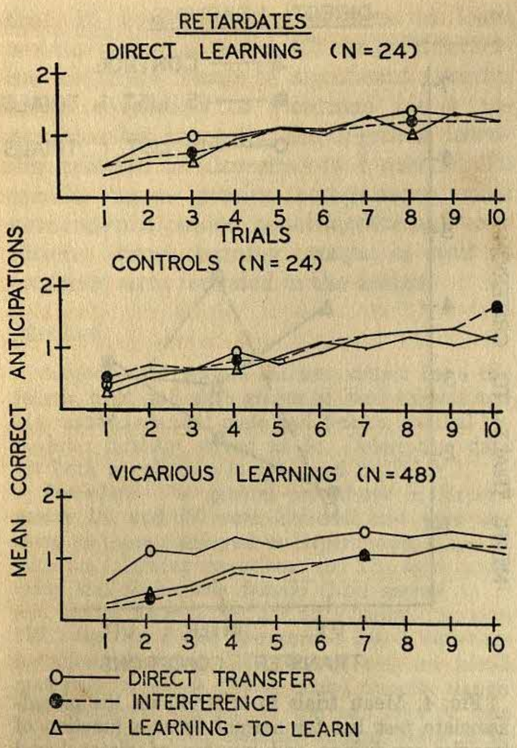


FIG. 5. Mean number of correct anticipations per transfer pair on Trials 1–10 of the paired-associate test list for retardate direct learning, vicarious learning, and control *S*s (Experiment IV).

variance on test list anticipation scores. Results indicated significant effects for Number of Vicarious Trials ($F = 7.43$, $df = 1/88$, $p < .01$), Transfer Conditions ($F = 19.51$, $df = 2/176$, $p < .001$), Trials ($F = 113.75$, $df = 4/352$, $p < .001$), Transfer Conditions × Trials ($F = 4.94$, $df = 8/704$, $p < .001$), and Instructions × Transfer Conditions × Trials ($F = 2.71$, $df = 8/704$, $p < .01$). Reliably more correct responses were made with 15 vicarious trials ($M = 1.52$) than with 5 trials ($M = 1.31$). The triple interaction is presented in the lower panels of Figure 3. As is evident, instructions to learn significantly facilitated performance on DT items and resulted in a greater (although insignificant) tendency for INT to occur on Trial 1 as compared with instructions to listen. In agreement with the trends noted in Experiment I, overall performance under both instructional conditions is somewhat below the performance of DL *S*s, the combined means for DT, INT, and LTL pairs being 1.59, 1.30, and 1.36, respectively.

A similar analysis on total omissions revealed that Transfer Conditions was the only significant source of variance ($F = 13.29$, $df = 2/176$, $p < .001$). The means for DT, INT, and LTL pairs were 1.20, 2.08, and 1.88, respectively, the latter two not differing significantly from each other. For total intralist intrusions significant effects were obtained for Number of Vicarious Trials ($F = 5.96$, $df = 1/88$, $p < .05$) and Transfer Conditions ($F = 5.12$, $df = 2/176$, $p < .01$). The means for 5 and 15 vicarious trials were 1.49 and .79, respectively. For DT, INT, and LTL pairs, the means were .83, 1.27, and 1.31.

Experiment IV. A 2 (direct learning versus controls) × 2 (number of List-1 trials) × 3 (transfer) × 10 (trials) mixed analysis of variance on performance over Trials 1–10 revealed that Trials ($F = 262.75$, $df = 9/396$, $p < .001$) was the only significant source of variance. Thus, contrary to results obtained with normal *S*s, retardates failed to show significant test list transfer as a result of directly learning an initially related list of PAs for either 5 or 15 trials. For DL *S*s, the overall means for DT, INT, and LTL pairs were 1.12, 1.00, and 1.01, respectively; the corresponding means for controls were .95, 1.00, and .85.

A 2 (listen versus learn) × 2 (number of vicarious trials) × 3 (transfer) × 10 (trials) mixed analysis of variance for VL groups indicated significant effects for Transfer Conditions ($F = 15.94$, $df = 2/88$, $p < .001$), Trials ($F = 29.80$, $df = 9/396$, $p < .001$), and Transfer Conditions × Trials ($F = 1.64$, $df = 18/792$, $p < .05$). As can be seen in Figure 5, more correct responses were made to DT pairs ($M = 1.23$) than to INT ($M = .88$) and LTL ($M = .88$) pairs. For comparison purposes, the performances of DL and control *S*s are also presented in this figure.

As in Experiment III, error analyses were consonant with correct response data. Comparing DL and control *S*s, analyses of variance on the first two blocks of five trials revealed that Blocks was the only

significant source of variance for both omissions ($F = 48.30$, $df = 1/44$, $p < .001$) and intralist intrusions ($F = 10.81$, $df = 1/44$, $p < .001$). In the omissions analysis for VL *S*s, Transfer Conditions ($F = 3.97$, $df = 2/88$, $p < .05$) and Blocks ($F = 43.33$, $df = 1/44$, $p < .001$) were statistically significant. The mean omissions for DT, INT, and LTL items were 1.98, 2.57, and 2.53, respectively. For intrusions, significant ratios were obtained for Transfer Conditions ($F = 8.30$, $df = 2/88$, $p < .001$) and Transfer Conditions × Blocks ($F = 4.55$, $df = 2/88$, $p < .05$). The latter reflects decreasing differences between DT, INT, and LTL items from Block 1 to Block 2. The overall means were 1.65, 2.30, and 2.57, respectively.

DISCUSSION

The major goals of this research were (*a*) to demonstrate the existence of vicarious PA learning in normals and retardates, (*b*) to examine the relative efficacy of direct versus vicarious procedures, and (*c*) to study some of the variables which may be of common importance in both these conditions of acquisition.

With reference to the first aim, the present results indicate that, following exposure to the errorless performance of a model, both normal children and retardates will learn the same PAs faster in transfer than will *S*s having no prior PA experience. The marginally significant trends for direct transfer pairs in Experiments I and II, as well as the reliable findings in Experiments III and IV, jointly support this conclusion. It is also clear, however, that instructional conditions during the observational trials are of primary importance in producing this effect for normals but not for the mentally retarded. Thus, for normal *S*s, instructions to learn the materials the model was learning produced reliable transfer in Experiment III, whereas listen instructions yielded nonsignificant and consistently lower levels of performance in Experiments I and III. Independent of how this instructional effect is conceptualized, it is important to note that these findings are consonant with results reported in the intentional-incidental learning literature and suggest some measure of correspondence between the two areas of investigation. That facilitation was obtained in Experiment IV independent of instructional conditions supports the notion that retardates are more outer-directed than normals (Zigler, 1966) and indicates that under conditions designed to facilitate orienting responses to the task-relevant cues, for example, exposure to the selective behavior of a social model, incidental-intentional learning differences fail to appear.

Both associative facilitation and interference effects have been reported for normal children (e.g., Norcross & Spiker, 1958) and retardates (e.g., Berkson & Cantor, 1960; Rieber, 1964) when items are learned directly in all stages of PA transfer paradigms. Under the present conditions of vicarious List-1 learning, however, no such effects were obtained. Herein, it should be noted that while a trend towards interference seems to occur on Trial 1 for intentional learning *S*s in Experiment III, (*a*) this trend was not statistically reliable, and (*b*) there was an operant level tendency for INT items to be somewhat more difficult than DT and LTL items in all groups, including controls.

Data relevant to the second aim of this research, that is, differential rates of learning under direct and vicarious conditions, appear to provide reasons for these discrepancies. For normal *S*s in Experiments I and III, there is a consistent trend for test performance to be higher when List-1 pairs are learned directly than when they are learned vicariously. With the exception of DT pairs in Experiment IV, the same trend was obtained for retardates in Experiments II and IV. If PA learning is now conceptualized as a two-stage process (Underwood & Schulz, 1960), then it is clear that conditions conducive to low levels of List-1 learning, for example, vicarious procedures, will primarily favor the response learning stage, whereas conditions producing higher levels of learning, for example, direct procedures, will tend to facilitate the associative learning stage. Thus, in the absence of appreciable second-

stage learning, neither associative facilitation nor interference would be expected. Indeed, with primarily response learning, the only items expected, and found, to show reliable transfer would be DT pairs. This interpretation is highly consistent with Martin's (1965) more extensive theoretical treatment of verbal paired-associate transfer and is consonant with a study recently reported by Winnick (1966) who failed to find differences between LTL and A-Br paradigms with a 50% List-1 learning criterion.

Additional support for this interpretation is provided by noting that reliable interference was obtained only in normal *S*s receiving 15 List-1 trials directly. An examination of performance on List 1 indicates that the mean number of correct responses per transfer pair for the five blocks of three trials is 1.24, 1.66, 1.82, 1.86, and 1.95, respectively. Further, 83% of these *S*s reached an overall criterion of at least five-sixths by the sixth trial. For retardate *S*s, the corresponding means for Blocks 1–5 are .75, 1.24, 1.38, 1.58, and 1.56, with 50% of the *S*s reaching the same criterion by the seventh trial and 68% by the fifteenth trial. For both populations, the means for DT, INT, and LTL pairs are similar over all trials. It is clear, therefore, that significant interference occurred under conditions of relatively marked first-list learning and that such conditions were neither present in the majority of direct learning retardate *S*s nor likely to be present in any *S*'s learning List 1 vicariously.

Seemingly contradicting the foregoing reasoning concerning the superiority of direct as against vicarious acquisition, VL *S*s in Experiment IV showed significant transfer to DT pairs whereas DL *S*s were no different from controls on any of the items. Since previous discussion must necessarily eliminate insignificant amounts of List-1 learning as primarily responsible, it is likely that these differences are attributable to conditions of testing. Specifically, the differences may be due to the "multiple rule" requirement of the mixed-list transfer design. This refers to the fact that with this design as many strategies are required of *S*s as there are transfer paradigms. It has been shown that relative to unmixed lists, performance with mixed lists is slightly poorer in a number of paradigms (Postman, 1966), including A-B, A-B (Slamecka, 1967). Presumably this is a result of interference produced by attempts to apply multiple rules to differing transfer pairs. Relative to VL *S*s, and assuming correspondingly greater amounts of associative learning on List 1, the somewhat depressed performance of DL *S*s on DT pairs would appear to agree with these findings. That the same decrement was not noted for DL *S*s in Experiment III, however, may be due to the fact that there was considerably greater List-1 learning for these *S*s and/or that normals are better able to overcome this source of interference than retardates.

The significant facilitation obtained under conditions of vicarious List-1 learning in Experiment III was independent of instructions concerning the scholastic competence of the model. In Experiments I and II, preferred peers selected by *S*s from their own classrooms or cottages to serve as models were equally ineffective in producing reliable transfer following vicarious List-1 learning. While other research has frequently found this variable to be of considerable importance in observational learning (Bandura & Walters, 1963), the present results suggest that more traditional verbal learning variables are likely to be of greater importance in vicarious paired-associate transfer.

In agreement with results reported by Otto (1961), reading level was significantly related to PA test performance in Experiment I: *S*s scoring high on a reading comprehension test learned the PA list significantly faster than *S*s scoring in either the middle or low ranges of the reading test. Inasmuch as reliable treatment effects were not obtained in this study, however, statements concerning possible differences in vicarious learning for these three groups must await further research.

REFERENCES

ANDERSON, R. C. Educational Psychology. *Annual Review of Psychology*, 1967, **18**, 129–164.

BANDURA, A., & MCDONALD, F. J. The influence of social reinforcement and the behavior of

models in shaping children's moral judgments. *Journal of Abnormal and Social Psychology,* 1963, **67,** 272–281.
BANDURA, A., & WALTERS, R. H. *Social learning and personality development.* New York: Holt, 1963.
BERGER, S. M. Conditioning through vicarious instigation. *Psychological Review,* 1962, **69,** 450–466.
BERKSON, G., & CANTOR, G. N. A study of mediation in mentally retarded and normal school children. *Journal of Educational Psychology,* 1960, **55,** 129–134.
DENNY, M. R. Research in learning and performance. In H. Stevens & Heber, R. (Eds.), *Mental retardation.* Chicago: University of Chicago Press, 1964.
DITRICHS, R., SIMON, S., & GREENE, B. Effect of vicarious scheduling on the verbal conditioning of hostility in children. *Journal of Personality and Social Psychology,* 1967, **6,** 71–78.
ELKONIN, D. B. The physiology of higher nervous activity and child psychology. In B. Simon (Ed.), *Psychology in the Soviet Union.* Stanford: Stanford University Press, 1957.
FLETCHER, H. J. Discrimination learning by retardates as a function of number of implicit response trials. *Psychonomic Science,* 1966, **4,** 159–160.
KANFER, F. H., & MARSTON, A. R. Human reinforcement: Vicarious and direct. *Journal of Experimental Psychology,* 1963, **65,** 292–296.
MARSTON, A. R. Variables in extinction following acquisition with vicarious reinforcement. *Journal of Experimental Psychology,* 1964, **68,** 312–315.
MARSTON, A. R. Determinants of the effects of vicarious reinforcement. *Journal of Experimental Psychology,* 1966, **71,** 550–558.
MARTIN, E. Transfer of verbal paired associates. *Psychological Review,* 1965, **72,** 327–343.
MCDAVID, J. W. Imitative behavior in preschool children. *Psychological Monographs,* 1959, **73**(16 Whole No. 486).
MCLAUGHLIN, B. "Intentional" and "incidental" learning in human subjects. *Psychological Bulletin,* 1965, **63,** 359–376.
NORCROSS, K. J., & SPIKER, C. C. Effects of mediated associations on transfer in paired-associate learning. *Journal of Experimental Psychology,* 1958, **55,** 129–134.
OTTO, W. The acquisition and retention of paired associates by good, average, and poor readers. *Journal of Educational Psychology,* 1961, **52,** 241–248.
PALERMO, D. S., & JENKINS, J. J. *Word association norms grade school through college.* Minneapolis: University of Minnesota Press, 1964.
PASCHKE, R. E., SIMON, S., & BELL, R. W. Vicarious discrimination learning in retardates. *Journal of Abnormal Psychology,* 1967, **72,** 536–542.
POSTMAN, L. Differences between unmixed and mixed transfer designs as a function of paradigm. *Journal of Verbal Learning and Verbal Behavior,* 1966, **5,** 240–248.
RIEBER, M. Verbal mediation in normal and retarded children. *American Journal of Mental Deficiency,* 1964, **68,** 634–641.
SIMON, S., DITRICHS, R., & JAMISON, N. Vicarious learning of common and uncommon associations in children. *Psychonomic Science,* 1965, **3,** 345–346.
SLAMECKA, N. J. Transfer with mixed and unmixed lists as a function of semantic relations. *Journal of Experimental Psychology,* 1967, **73,** 405–410.
UNDERWOOD, B. J., & SCHULZ, R. W. *Meaningfulness and verbal learning.* Philadelphia: Lippincott, 1960.
VAN WAGENEN, R. K., & TRAVERS, R. M. W. Learning under conditions of direct and vicarious reinforcement. *Journal of Educational Psychology,* 1963, **54,** 356–362.
WINNICK, W. A. Effect of instructional set and amount of first learning on negative transfer. *Journal of Experimental Psychology,* 1966, **71,** 920–923.
ZEAMAN, D., & HOUSE, B. J. The role of attention in retardate discrimination learning. In N. R. Ellis (Ed.), *Handbook of mental deficiency.* New York: McGraw-Hill, 1963.
ZIGLER, E. Research on personality structure in the retardate. In N. R. Ellis (Ed.), *International review of research in mental retardation.* Vol. I. New York: Academic Press, 1966.

(Received August 10, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 5, 350-354

EXTRA-SCOPE TRANSFER IN LEARNING MATHEMATICAL STRATEGIES[1]

JOSEPH M. SCANDURA AND JOHN H. DURNIN
University of Pennsylvania

The notion of a restricted rule or strategy was introduced. It was hypothesized that extra-scope transfer depends on the extent to which a statement of strategy may be viewed as a restriction of a more general strategy. 66 high school *S*s were taught a restricted statement (S′, SG′, or G′) of 1 of 3 strategies of varying generality, S (= S′) < SG (SG′) < G (G′). 22 *S*s served as a control (C). All *S*s were tested on 6 problems, the 1st 2 within the scope of the most specific strategy (S), the 2nd 2 within the scope of only the more general strategies (SG and G), and the last 2 only within the scope of strategy G. Statements S′, SG′, and G′ were directly applicable only to the 1st 2 problems. Groups SG′ and G′ evidenced extra-scope transfer. Groups S′ and C did not. In addition, performance on the 2nd problem of each pair was contingent on performance on the corresponding 1st problems indicating that "what is learned" may be determined by performance on single test items and used to predict performance on additional similar-scope problems. Suggestions were made for future research.

Scandura, Woodward, and Lee (1967) demonstrated that performance on transfer tasks is generally in accord with the logically determined scope of rule and strategy statements.[2] In each of two experiments, *S*s were presented with one of three statements of rules (or strategies) of varying generality and were tested on three problems. The first problem was within the scope of all three rules; the second, within the scope of only the two more general rules; and the third, within the scope of only the most general rule. In most instances, there was essentially no difference in the level of performance on the within-scope problems and no extra-scope transfer (to problems to which the rule did not directly apply).

There was one glaring exception involving extra-scope transfer. In Experiment II, *S*s given the statement, "50 × 50," which was directly applicable only to Problem 1, performed equally as well on Extra-Scope Problem 2 as did those *S*s given the statement, "*n* × *n*," where the dimension (i.e., variable) *n* was allowed to vary over the positive integers. This result obtained even though "*n* × *n*" was directly applicable to both Problems 1 and 2. While the study itself was inadequate to specify the source of this transfer, a post hoc analysis of the experimental treatments indicated that "50 × 50" was the only rule statement included in the study which was in some sense a restriction of a more general rule or strategy. The statement, "50 × 50," could be obtained from the more general statement "*n* × *n*," by replacing the response determining dimension, *n*, by the value 50. More generally, it would appear that a restricted statement may be viewed as one obtained by replacing response-determining dimensions (see Scandura, 1966, 1967a, 1968a) in the statement of a general rule or strategy with the specific values of a particular instance. The authors, therefore, conjectured that a restricted rule statement might well provide a basis for generalization to all problems within the scope of the corresponding *un*restricted rule. The pri-

[1] The participation of the second author was made possible by a United States Office of Education Graduate Training Grant to the University of Pennsylvania in Mathematics Education Research. The authors are indebted to I. R. Klingsburg, head of the mathematics department, and the participating students at the West Philadelphia High School for their generous cooperation in making this study possible.

[2] The terms "rule" and "strategy" are used synonymously throughout this paper. While "rule" is the preferred technical term (e.g., Scandura, 1968a), "strategy" better connotes more complex multiphased rules of the sort used in this study.

mary purpose of this study was to test this hypothesis.

A secondary purpose was to obtain further information on the "consistency" hypothesis. Under certain conditions, it has been found that transfer to one instance of a rule almost invariably implies transfer to other instances of the rule (Scandura, 1966, 1967a, 1967b, 1968b; Scandura et al., 1967). As was the case with extra-scope transfer, however, one exception to the consistency hypothesis was obtained in the study by Scandura et al. (1967). The level of performance on one within-scope problem was considerably below that on the others. Whereas the response determining values of the homogeneous problems differed along a single dimension, the exceptional within-scope problem differed along a second dimension as well. Taking this observation into account, a modified form of the consistency hypothesis was advanced. It was hypothesized that if transfer to one problem indicates that a particular rule or strategy (e.g., "50 × 50") has been generalized along one or more familiar dimensions (e.g., to "$n \times n$") then transfer to additional problems along the same dimensions (and within the scope of a less restrictive rule) should also be expected.

METHOD

Material

The material was based on a variant of the number game, "NIM." In this variant, two players alternately select numbers from a specified set of consecutive integers (including 1) and keep a running sum. The winner is the one who picks the last number in a series having a predetermined sum. Each such game can be characterized by an ordered pair (n, m) where the corresponding value of n is the largest integer in the selection set and the value of m is the predetermined sum (n and m refer to dimensions over which NIM may vary). If the set consists of the integers 1–6 and the sum is 31, the players alternately select numbers 1–6 until the cumulative sum is either 31 or above (in which case no one wins).

Scandura et al. (1967) presented statements of three general rules by which the person making the first selection can always win. The specific (S) rule is sufficient for winning only (6, 31) games and was stated:

> In order to win the game you should make 3 your first selection. Then you should make selections so that the sums corresponding to your selections differ by 7.

The specific-general (SG) rule, an unidimensional strategy, is applicable to any game of the form (6, m) and was stated:

> In order to win the game, the appropriate first selection is determined by dividing the desired sum by 7. The remainder of this division is precisely the selection which should be made first....

The general (G) rule, a bidimensional strategy, is applicable to any (n, m) game, where both n and m are allowed to vary, and was stated:

> In order to win the game the appropriate first selection is determined by adding one to the largest number in the set from which the selections must come and dividing the desired sum by this result. The remainder of this division is precisely the selection that should be made first. Then you should make selections so that the sums corresponding to your selections differ by one greater than the largest number in the set from which the selections must come.

Statements of *restrictions* of these strategies, which are applicable only to (6, 31) games, were constructed for use in this experiment. Rule S′ was essentially identical to rule S and was stated in the same way. Rule SG′ was a restriction of unidimensional strategy, SG, in the sense that SG was restricted to one value (i.e., 31) of the desired-sum (m) dimension. Rule SG′ was stated:

> The appropriate first selection is determined by dividing 31 by 7. The remainder 3 should be your first selection....

Rule G′ was a restriction of bidimensional strategy, G, in that G was restricted to one value (i.e., 31 and 6, respectively) of the desired-sum (m) and size-of-selection-set (n) dimensions. Rule G′ was stated:

> The appropriate first selection is determined by adding one to six, (1 + 6), and dividing 31 by this result. The remainder 3 of this division is the selection which should be made first.... It is important to notice that 7 = 6 + 1.

The materials were reproduced by mimeograph and were combined into nine different 8½ × 11 inch booklets—an introduction, four treatments, and four tests. The introduction consisted of 4 pages and was designed to insure that *S*s knew the objective of and how to play NIM. Page 1 indicated that the experimental results would be made available to *S*s, asked that they not divulge information about the experiment to others who might participate, and explained the nature of the (6, 31) game. Page 2 consisted of one completed (6, 31) example and a (6, 31) practice game in which *S* was required to compute the run-

ning sums in accordance with a specified sequence of selections. Knowledge of results was given on page 3 along with another (6, 31) practice game with the result of the latter given on page 4. Nothing was said in the introduction about game-winning rules, but it was mentioned that there are many variations of NIM.

Three of the four treatment booklets included one of the restricted statements (S′, SG′, and G′) on page 1 together with a common (6, 31) game which was provided for practice. The solution to this (6, 31) game was given on page 2 and *S*s were instructed to replay the same game, on page 3 after correcting any previous errors. In this common (6, 31) game, the running sums were $\underline{3}$, 5, $\underline{10}$, 11, $\underline{17}$, 20, $\underline{24}$, 25, $\underline{31}$. The fourth booklet served as a control. It consisted solely of the common (6, 31) example with no statement of a game-winning rule. Nonetheless, by remembering those sums which corresponded to the winning selections (i.e., 3, 10, 17, 24, 31), an *S* might conceivably win any new (6, 31) game.

The four test booklets corresponded to the four treatment booklets. Page 1 was common to all test booklets and explained how to use the booklet. Each of the successive pages (2–7) included one common test game together with that game-winning statement (S′, SG′, or G′) associated with the corresponding treatment booklet. This procedure was followed to help eliminate errors due to recall. The "opponent's" selections were printed in the booklet and *S* was instructed to make his selections and to compute the running sums. The first two problems, 1A and 1B, were (6, 31) games. Problems 2A and 2B were (6, m) games which differed along the desired-sum dimension, with $m = 25$ and $m = 29$, respectively. Problems 3A and 3B were (n, m) games, which differed along both the desired-sum and size-of-selection-set dimensions, with $n = 5$, $m = 26$ and $n = 7$, $m = 33$, respectively.

Subjects, Design, and Procedure

The *S*s were 88 West Philadelphia High School students enrolled in an academic mathematics program. They were randomly assigned to three experimental groups (S′, SG′, G′) and a control (C) so that each group included 22 *S*s.

Each *S* completed the introductory booklet, one of the four treatment booklets, and the corresponding test booklet, in that order. The *S* was told to read the material carefully. The experiment was self-paced and with only a few exceptions *S*s completed the experiment well within the time limit of 40 minutes.

The criterion measure was use of the appropriate pattern (AP). The *S* was given credit for using the AP if he won the game and employed an appropriate game-winning strategy. All of the tests conducted were applied to 2 × 2 contingency tables. When the measures were independent, the exact Fisher-Yates formula was used (Finney, 1948); when correlated, a different nonparametric test, based on χ^2, was used (McNemar, 1955, pp. 358–359). One-tailed tests were used in conjunction with the stated hypotheses with an alpha level of .05.

RESULTS AND DISCUSSION

Table 1 shows that restricted rule statements may provide an adequate basis for generalization. Statements of unidimensional and bidimensional strategies, even when restricted to particular values of these dimensions, may result in transfer to new problems which differ from the training problem (e.g., common example) along these same dimensions. The three experimental groups performed at essentially the same level on problems 1A and 1B, but there were 12 *S*s in Groups SG′ and G′, as compared to *none* in Group S′, who were successful on problems 2A and 2B. This difference was significant at the .01 level.

A cursory review of the literature suggests that the transfer observed in a number of other studies may also have involved generalizing along one or more dimensions of a restricted rule statement. Maier (1945), for example, found that providing *S* with a problem-solving strategy, as it applied to one problem (i.e., with a restricted statement), improved the level of performance on a second problem (which was presumably within the scope of a more general strategy). Some such generalization mechanism may also be involved in what some investigators have called "remote transfer." Thus, in a recent study,

TABLE 1

NUMBER OF APPROPRIATE PATTERNS

Group	N	Problem					
		1A (6, 31)	1B (6, 31)	2A (6, 25)	2B (6, 29)	3A (5, 26)	3B (7, 35)
C	22	0	0	0	0	0	0
S′	22	16	16	0	0	0	0
SG′	22	19	20	7	7	2	2
G′	22	18	18	5	5	0	2

Note.—Abbreviated: C = control, S′ = restricted specific, SG′ = restricted specific-general, G′ = restricted general.

Wittrock's (1967) nonreplacement-strategy group was presented with a restriction of a general strategy which was applicable to his remote transfer items. Apparently, what these *S*s actually learned (i.e., discovered) was the more general strategy.[3]

The performance of the G′ *S*s, however, suggests that transfer cannot necessarily be expected to all problems within the scope of the rule from which a restricted statement is derived. Of the five *S*s in Group G′ who were successful on problems 2A and 2B, none was successful on problem 3A and only two, on problem 3B. These differences between problems 2A and 2B and problems 3A and 3B suggest that the level of performance on transfer problems may depend on the particular dimension(s) involved. Problems 2A and 2B required that the G′ statement be generalized only along the desired-sum dimension whereas problems 3A and 3B required generalization along the size-of-selection-set dimension as well. Apparently, the G′ *S*s were more capable of making the former generalization than the latter.

The authors also feel obliged[4] to comment on the fact that two SG′ *S*s generalized beyond the scope of rule SG to problems 3A and 3B. These SG′ *S*s were apparently as able to generalize along the size-of-selection-set dimension as were the two G′ *S*s who were successful on problem 3B. Thus, the statement cue, "7," in statement SG′ was equally as helpful as the cue, "6 + 1," in statement G′ even though "6" in the latter cue corresponded directly to the number of integers in the selection set. (The former cue, "7," was one larger.) The S′ *S*s, on the other hand, seemed uniformly unable to generalize along either dimension. To do so, they would have had to have observed that the desired sum, 31, when divided by the constant difference, 7, leaves a remainder of 3 (the first selection).

These observations suggest that the ease with which response-determining properties of an illustrative (training) problem can be related to the corresponding response-determining value (cue) in a restricted statement may have an important effect on the extent of transfer. A pilot study conducted with 20 highly motivated and mathematically oriented doctoral students at the University of Pennsylvania tends to provide further support for this interpretation. All of the SG′ and G′ *S*s and four out of five of the S′ *S*s were able to generalize to problems 2A, 2B, 3A, and 3B. Clearly, the ease with which a correspondence can be determined between the determining properties of an illustrative problem and statement cues depends on individual differences as well as on the nature of the cue. A major task of future research will be to determine what the important individual differences are.

To test the consistency hypothesis, those *S*s who used the AP on problems 1A, 2A, and 3A and those who did not (non-AP users) were compared as to AP use on problems 1B, 2B, and 3B, respectively. There were significantly more AP users on problem 1A who were AP users on problem 1B than was the case for non-AP users on problem 1A ($p < .001$). The same relationship held for problems 2A and 2B ($p < .001$) and problems 3A and 3B ($p < .001$), respectively. There were only 4 cases out of a total of 131 in which a non-AP user (in Groups S′, SG′, and G′) on an A problem

[3] Many psychologists feel that "what is learned" is excess theoretical baggage since the notion must invariably be defined in terms of transfer. While admitting the ultimate necessity of operational definition, the authors take the position that "what is learned" is a useful construct. In particular, performance on two test items (one training and one transfer) can often be used to identify "what is learned" by individual *S*s, thereby making it possible to predict their performance on additional transfer items. This latter assertion is well exemplified by the present consistency data.

[4] A program of ongoing research by the first author and his collaborators is aimed at uncovering laws of mathematical learning and behavior which hold in a deterministic (or near-deterministic) sense. Thus, when exceptions occur, even where the effects are not "statistically reliable," they are viewed as facts to be explained and not probabilistic deviations which may be safely ignored. Although both the behaviors in question and the methods of approach differ greatly, the authors' research objectives are quite similar to those adopted long ago by Skinner and his followers—to uncover idiographic laws.

became an AP user on the corresponding B problem. There was only 1 case (out of 67) where an AP user on an A problem was *not* an AP user on the corresponding B problem.

These results suggest that if transfer obtains on one new problem, which differs (from the training problem) along one or more dimensions, then transfer may be expected to other problems which differ along these same dimensions. Of course, the boundary conditions for this assertion still need to be determined. At the very least, it would seem that the dimension(s) in question would have to be familiar to *S*s (but just what this familiarity entails is not entirely clear).

REFERENCES

FINNEY, D. J. The Fisher-Yates test of significance in 2 × 2 contingency tables. *Biometrika,* 1948, **35,** 145–156.

McNEMAR, Q. *Psychological statistics.* (2nd ed.) New York: Wiley, 1955.

MAIER, N. R. F. Reasoning in humans. III. The mechanisms of equivalent stimuli and of reasoning. *Journal of Experimental Psychology,* 1945, **35,** 349–360.

SCANDURA, J. M. Precision in research on mathematics learning; the emerging field of psychomathematics. *Journal of Research in Science Teaching,* 1966, **4,** 253–274.

SCANDURA, J. M. The basic unit in meaningful learning—association or principle? *School Review,* 1967, **75,** 329–341. (a)

SCANDURA, J. M. Learning verbal and symbolic statements of mathematical rules. *Journal of Educational Psychology,* 1967, **58,** 356–364. (b)

SCANDURA, J. M. Using the rule to formulate research on meaningful learning. I. A set-function language. *Acta Psychologica,* 1968, in press. (a)

SCANDURA, J. M. Using the rule to formulate research on meaningful learning. II. Empirical research. *Acta Psychologica,* 1968, in press. (b)

SCANDURA, J. M. Using the rule to formulate research on meaningful learning. III. Analyses and theoretical direction. *Acta Psychologica,* 1968, in press. (c)

SCANDURA, J. M., WOODWARD, E., & LEE, F. Rule generality and consistency in mathematics learning. *American Educational Research Journal,* 1967, **4,** 303–319.

WITTROCK, M. C. Replacement and nonreplacement strategies in children's problem solving. *Journal of Educational Psychology,* 1967, **58,** 69–74.

(Received September 18, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 5, 355–361

SOME FACTORS IN CHILDREN'S LEARNING AND RETENTION OF CONCRETE RULES[1]

ROBERT M. GAGNÉ AND VIRGINIA K. WIEGAND
University of California, Berkeley

An experiment investigated effects of (*a*) number of rules, (*b*) immediate vs. 3-day recall, (*c*) verbal vs. verbal plus pictorial cueing, and (*d*) IQ, on the learning and retention of concrete rules. Each rule was composed of a highly pronounceable CVC as the name of a thing (a drawn shape), and an action (such as, "underline it"). 96 4th-grade children were the experimental *S*s. Following sessions providing prelearning on thing concepts and review of action concepts, different groups of children learned 3, 5, 7, or 9 rules from printed booklets, reading and recording each rule once. Virtually complete retention was obtained for 3 and 5 rules when measured immediately, significantly less for 7 and 9. Significant effects were not found for IQ or cueing method. After 3 days, the amount of retention was about 20% under all conditions.

The learning and retention of ideas have usually been studied with the use of prose materials (e.g., Briggs & Reed, 1943; Cofer, 1941; Hall, 1955; King & Russell, 1966). In such investigations, the learner typically first listens to or reads a prose passage for a specified number of trials, and later is asked to recall or recognize certain selected ideas embodied in the passage. Such an approach generally assumes that what is learned and retained will be influenced by the meaningful context within which the particular ideas to be recalled are imbedded. The context may be presumed to provide cues, as appears to be true with paired associates (McGovern, 1964); it may be a source of mediators (Davidson, 1964); it may provide "organizers" of the sort studied by Ausubel (1962); or it may have still other effects, as yet unidentified.

In contrast to this context-imbedded approach, the present authors were interested in taking an analytic view of the learning of ideas, one which at the outset attempts to separate the effects of context variables from others. This investigation was begun with a specification of the concept of "idea" that was as definite as possible. In pursuit of the latter goal, research of previous writers and theorists provided scant information. Accordingly, in order to proceed with this account, the authors had to propose some of their own definitions. These are tentative, subject to further development, and are given here for the purpose of aiding communication.

For the concept of "idea" itself, the authors refer to Gagné (1965), who considers an idea to be a principle, or rule. The latter is a capability inferred as having been learned, which enables the individual to demonstrate a sequence of concepts such as "If A, then B." The concepts which make up a rule are (simpler) capabilities, evidenced in overt behavior by the identifying of classes of objects or events. Included as concepts are classes of the individual's own actions. A rule may, of course, include a rather large number of concepts in its sequence. However, the simplest rule in a formal sense is one that contains only two concepts. A common variety is a rule combining a "thing concept" with an "action concept," as in "birds fly," "water wets," and "If green, go."

A rule may be concrete or abstract, depending upon whether its component concepts are concrete or abstract. If the two components of a rule may be represented by a concrete noun and a verb describing

[1] The research reported herein was performed pursuant to a contract with the United States Department of Health, Education, and Welfare, Office of Education. Thanks are due to William Flynn, principal, and to the fourth-grade teachers of the Buena Vista School, Walnut Creek, California—Sally Collins, Carol Richman, Arthur Sloane, and Jean Wagner.

overt action, the rule is concrete. A partially abstract rule combines a concept represented by a concrete noun with a concept represented by an abstract verb, as exemplified by "women suffer" or "snakes live." The reverse combination, abstract noun–concrete verb, may also be called partially abstract. There are also fully abstract rules, like "power corrupts." The authors believe these distinctions are of some importance to the study of rules. However, no more extensive discussion of them will be undertaken here; they are mentioned in order to convey the meaning of "concrete rule."

The present study had the purpose of investigating some factors with potential influence on the learning and retention of concrete rules, of the sort sometimes called facts. The major variable of interest was the number of such rules to be learned in a single set, when presented one after another. Another variable was the manner of cueing used in eliciting the rule during a test of retention. Both immediate and delayed (3-day) retention were measured.

It is of some interest to note that the amount of retention of learned ideas varies over a wide range, as reported in different studies. It is also true that variations in procedure are many, making comparisons among studies difficult or impossible. An early study by Yoakam (1921), for example, found retention of a prose passage read a single time to be from 48% to 80% as complete after 20 days as retention measured immediately, in different groups of children. In the study of Dietze and Jones (1931), a single reading of a 1,000-word article yielded scores ranging from 44% to 84% for immediate retention, and from 26% to 42% after 30 days. Newman (1939) reports 87% retention after 8 hours of sleep, 86% after 8 hours of waking, of 12 main ideas contained in a 300-word story. Comparable scores were 47% after sleep and 23% after waking for 12 nonessential ideas. In Cofer's (1941) study of the learning of prose passages, no evidence of retention was found after 9 months, as measured by relearning scores, in *S*s who learned the material for logical (as opposed to verbatim) reproduction.

Rationale for an Analytic Study

It is assumed that a single concrete rule, whose component concepts are readily available to recall, is learned in a single trial. If one then undertakes to learn two or more such rules in a row, the possibility exists that they will not exhibit perfect immediate retention. How many such rules can be learned (and retained) when presented one after another? What is the effect of the variable of number of concrete rules on the retention of such rules as indicated both by a test immediately following learning and by a test given 3 days later? These are the questions addressed in this study.

Specifically, fourth-grade children who had previously learned thing concepts, and recalled some familiar action concepts, were given concrete rules to learn. Each rule combined a thing (a drawn shape) with an action (such as underlining the shape). In designing the thing and action concepts making up each rule, the factor of familiarity was considered. Since retention of the rules was to be cued by verbal names, it was desired to keep them equally unfamiliar, and thus to avoid the possible differential effects of mediating verbal associations. Accordingly, nonword names were learned for the things. For action concepts, however, since these were to be recalled, no necessity was seen to insure that they were unfamiliar. Instead, they were chosen to be highly familiar and highly overlearned.

The children learned these rules from the pages of booklets, each page of which gave a printed statement of the rule and then asked the learner to draw it. Different groups of children learned different numbers of rules: three, five, seven, or nine. Retention of the rules was then measured immediately; and in separate groups, after 3 days. For the purpose of investigating cues to recall, half of the retention booklets cued the rules by means of only a verbal name of the thing concept, half by the verbal name plus a drawing of the thing concept.

METHOD

Subjects

The learners in the study were 96 fourth-grade children in an elementary school located in a primarily middle-class suburban community. With the cooperation of their four teachers, the experiment was conducted as an exercise in science instruction, in which the children were informed participants. The study began with 104 children. The rules of random assignment were followed throughout, with the exception that two students were excluded because they did not complete the prelearning. With absences and other contingencies taken care of, there was a total N of 96 for the study.

Materials

Prelearning of thing concepts. Four different study sheets were prepared to accomplish the prelearning of thing concepts. Each sheet showed a thing (shape) with its printed name (nonsense syllable) underneath it in nine boxes having a scattered arrangement on the page. Ten different test sheets were also made up, containing nine boxes running vertically, each with a printed concept name and a space for the shape to be drawn.

The concept names, typed in capital letters, were nine consonant-vowel-consonant (CVC) syllables with relatively high pronounceability values, averaging 2.49, as given by Underwood and Schulz (1960). No consonants were used in common in either the first or third position, and vowels were balanced in frequency as nearly as possible (BOT, FAC, JUM, LAB, NOP, REL, SUD, TIS, ZIN). The figures were chosen to be distinctive and easy to draw; they included such common figures as a square, a triangle, a heart, and a crescent. In successive study sheets, the figures were shown in four different orders from the top to the bottom of the page. In addition, considerable variation was introduced into nonrelevant features such as size and thickness of outline, without changing the essential shape. A similar plan was followed for test sheets.

Review of action concepts. A booklet was used to present nine different actions. Each action was first shown accompanied by printed instructions which were read by E (i.e., "There is a line under the shape."). The child then executed the action with a second figure of the same kind on the same page, and again with another figure on the next page. The figures employed in this booklet were different from those used in learning. The nine actions represented were: line under, tail on, circle around, question mark after, dot in, check before, hat over, X through, and legs on. The booklet contained a final test for all nine actions, presented four on one page and five on the next, with printed instructions ("Draw a circle around the shape.") for each action.

Rule learning. The booklets for rule learning were designed to be used by each child in individual learning. Each page first identified a thing concept ("This is a NOP."). Then the statement of the rule was made: "Rule: A NOP has a circle around it." The rule was then illustrated. Finally, the printed instructions said, "Draw the rule for a NOP," and a blank space was provided where this was to be done. To meet design requirements, a fourth of the booklets presented three rules; a fourth, five; a fourth, seven; and a fourth, nine.

Each thing concept, and each thing-action combination, was made to have an approximately equal frequency of occurrence in booklets for each of the treatment groups pertaining to number of rules (three, five, seven, and nine). Within this constraint, there was random assignment of rule booklets among Ss within each of these groups.

Filler activities. Several booklets were prepared containing filler activities, to be used when each learner finished his learning booklet. These were necessary in view of the fact that the number of rules presented in learning was as few as three for some Ss, and as many as nine for others. The filler booklets contained both geometrical and numerical puzzles.

Retention. Booklets used to measure retention contained 1 page devoted to each rule. The four varieties of booklets were 3, 5, 7, and 9 pages long, to correspond with the number of rules learned by different groups of Ss. Each page contained a printed statement of the form, "Draw the rule for a NOP." To permit the investigation of the cueing variable, half of the booklets included a drawn figure of each thing concept just below its name.

Design

The Ss were first divided into six subgroups representing six levels of IQ, as measured by traditional group test scores. In these Ss, this variable had a median of 110, a range of 86–149, and a roughly normal distribution. The Ss within each IQ level were then assigned randomly to each of the 16 treatment conditions of the experiment. At the beginning of the study, it was possible to assign seven or eight to each condition. After attritional factors had taken their toll, the assignment of Ss to conditions ended up as shown in Table 1, achieved by the final discarding of data from only three Ss, chosen by a random process.

TABLE 1

NUMBER OF SUBJECTS ASSIGNED TO TREATMENT CELLS REPRESENTING THE EXPERIMENTAL CONDITIONS

Retention condition	Cueing condition	Number of rules			
		3	5	7	9
Immediate	Verbal	6	6	6	6
	Verbal + Pictorial	6	6	6	6
Delayed	Verbal	6	6	6	6
	Verbal + Pictorial	6	6	6	6

It will be seen that the design is a four factorial one, making possible the determination of the effects of four different variables: (*a*) number of rules learned; (*b*) immediate versus 3-day retention; (*c*) verbal versus verbal plus pictorial cueing of recall; and (*d*) IQ level.

Procedure

The schedule for the study was as follows: Days 1–4, prelearning of names of thing concepts; Day 5, review of action concepts; Day 6, learning of rules, followed by retention test for half of the total group; Day 7 (preceded by an interval of 72 hours), retention test for the other half of the group. Each day's session occupied approximately a ½-hour period.

The prelearning of the names of thing concepts (shapes) began with a session in which *E* gave a general introduction to the study. Following this, *E* drew each shape on the blackboard and placed beside it its printed name. When all nine shapes had been introduced in this way, study sheets were passed out to all the children. They were asked to learn the names of all nine shapes. When a child thought he knew them all, he asked for a test sheet, and turned in his study sheet. When this in turn was turned in, he received another study sheet, followed by a second test sheet. Two study-test trials of this sort were conducted during each session for 4 days. In addition to the introductory trial, each *S* thus had a total of eight such trials. On the second test given on the fourth day, only two of the children did not identify all of the names correctly; data from these children were not used in the experiment. The criterion of the prelearning of thing concepts was thus a conjunctive one: completing eight study trials and identifying all nine concepts correctly on the final test.

On the day following the prelearning of thing concepts, the review and test session was conducted for the nine action concepts. Following a brief introduction, these trials were administered by means of the booklet previously described. All the children knew the action concepts after the first trial (and most probably before it).

Whereas prelearning and the action concept review were conducted in the four regular classrooms, the children were assembled in one large room for the rule-learning and retention sessions. Their places at the tables in this room were determined before their assembly, and booklets pertaining to the respective subgroups of the experiment were distributed at these places. In the learning session, the children were first instructed to work through the booklets page by page, and when finished, to go on to the second booklet. They were also asked to record the time of starting each booklet. In this manner, once the trials for rule learning were finished, the learners went on to the other activities contained in the filler booklets, regardless of whether they had studied three, five, seven, or nine rules. The session was brought to an end when it was observed that all children having booklets containing nine rules had finished. Each rule the child encountered was first presented in printed form and illustrated, and then drawn by him on the lower part of the same page. The mean time per rule (each rule being composed of these two steps of observing and recording) was found to be .33 minutes, in the three-rule group; .23 minutes in the five-rule group; .22 minutes in the seven-rule group; and .21 minutes in the nine-rule group.

For half of the total group, distributed equally in relation to the other variables of the experiment, the second booklet was a test of immediate retention, whereas for the other half, the second booklet contained filler activities. Retention booklets began with no preliminary instructions, simply containing a statement like "Draw the rule for a ZIN" on each page, as previously described. Those who finished these booklets before others went on to a third booklet of filler activities.

Delayed retention was measured with the other half of the total group on a Monday, following a weekend, 3 days after the rule-learning session. In its essentials, the same procedure was followed as for immediate retention. It may be noted that these children had not been told on Friday that they would be asked to recall the rules they had learned.

RESULTS

The raw data of the experiment were number of rules recalled, under the various conditions described in Table 1, and for six IQ levels within each of these conditions. An examination was also made of total overt errors, as well as of various categories of overt errors. Insofar as sufficient data existed, trends in errors appeared similar to those of correct responses; therefore, no more detailed analysis was undertaken.

The data first were converted to percentage remembered. An analysis of variance of these data revealed a significant effect of the immediate versus delayed retention variable ($F = 147$, $df = 1/15$, $p < .01$). Nonsignificant differences were found for the variable of IQ; and also for the variable of verbal versus verbal-plus-pictorial cueing. The analysis did not reveal differences for the variable of number of rules learned; for the data as a whole, the percentage of retention was not significantly different whether three, five, seven, or nine rules had been learned ($F = 2.3$, $df = 3/15$, $p > .05$). In addition, no significant interactions were found.

Cueing effects

The extra cue to recall of the correct rule, namely, the provision of a picture of the thing concept rather than only its name, is not shown by these results to have aided recall of the rule ($F < 1$, $df = 1/15$, $p > .05$). Thus, the group of children who were cued to recall only by the name of the thing concept contained in it, had no more difficulty in recalling the rule than did children who were cued by a picture plus the name. These results may also be considered as evidence that the children were able to use the name effectively, and that their performance in recalling rules was not affected by inability to recall the names per se.

Number of rules learned

The mean percentage recalled is shown (for the verbal and verbal-plus-pictorial groups combined) in Figure 1, for three, five, seven, and nine rules, under conditions of immediate and 3-day retention. It appears that three and five concrete rules are learned virtually intact, when recall is measured immediately. This near-perfect performance appears to drop off, however, when seven rules are learned, and to undergo a further drop for nine rules. The special circumstance to be considered here is that, as compared with these mean values, the performance of the group which learned five rules is 100%, with an *SD* of 0. It would seem legitimate to pose the question as to whether the means for the seven-rule and the nine-rule group are significantly different from a perfect performance. The usual statistical tests, however, are not designed to deal with measures having zero variance. Accordingly, the following test was applied.

The standard error of the mean for the group which learned seven rules was 7.6, yielding a confidence interval at the 99% level extending to the score 98.9. The mean of 79.8 for this group is thus significantly different from 100, the score attained by the group which learned five rules. Similarly, for the group which learned nine rules, the upper limit of the confidence interval is 77.9, and by similar reasoning

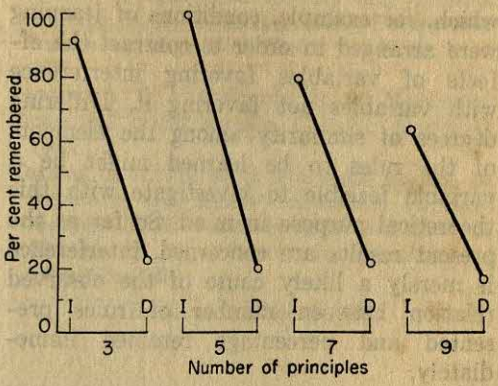


FIG. 1. Percentage of principles remembered immediately following learning (I) and after 3 days (D), by groups who were given different numbers of principles (three, five, seven, and nine) to learn.

it may be said that the mean of this group (63.9) differs significantly from the perfect performance of the five-rule group. A *t* test of the difference between the immediate retention scores of the seven-rule and nine-rule groups indicated nonsignificance at the .05 level.

As for percentage recalled after 3 days, this tends to be around 20%, and shows only slight and insignificant change depending upon the number of rules learned.

DISCUSSION

There appear to be some useful facts in the results of this experiment. When both thing concepts and action concepts are previously well learned, fourth-grade children can learn up to five concrete rules by reading and recording each once, one after another, when they are studied at an average rate of one every 15 seconds. When retention of seven and nine such rules is measured immediately, however, losses of the order of 20% and 36%, respectively, are found.

The dropping off of immediate retention scores under these conditions would be expected to occur as a result of interference (cf. Keppel, 1968), both proactive and retroactive. This experiment, of course, provides no direct confirmation that interference, rather than some other factor, is at work. Such a conclusion would need to be based upon an additional study in

which, for example, conditions of learning were arranged in order to contrast the effects of variables favoring interference with variables not favoring it. Differing degrees of similarity among the elements of the rules to be learned might be a variable feasible to investigate with this theoretical purpose in mind. So far as the present results are concerned, interference is merely a likely cause of the observed relation between number of rules presented and percentage retained immediately.

The results also suggest another kind of possible relationship with theory. Postman (1964) points out the relationship of the immediate memory span to the initial events of memory for items to be reproduced serially. Jensen (1964) found a high degree of relationship between measures of immediate memory span and the learning of serial lists of verbal items. Although the rules of the present experiment were not to be learned in serial order, they were presented serially. It may be of some significance that the falling off of immediate retention scores occurs in the interval (five–seven items) which has often been measured as the limit of memory span in children of the age used in this experiment.

As for the amount retained after a 3-day delay, it is most interesting that this was no more than 20%. It should be borne in mind that several obvious factors favor this relatively low degree of retention. First, the children were given no particular incentive to remember, and it therefore seems likely that they undertook little rehersal during the period intervening between learning and recall. Second, the facts they were asked to learn and recall were isolated from each other and from any other meaningful context (except the most general one that they were learning rules containing new names for things). Under these circumstances, the 80% loss in retention appears to be the order of what may be expected. However, it is a high degree of loss when compared to that obtained in studies in which the ideas are imbedded in a context (e.g., Yoakam, 1921; Dietze & Jones, 1931; English, Welborn, & Killian, 1934).

The apparent contrast between the remembering of isolated versus context-imbedded ideas would appear to suggest one promising direction for additional study. On the one hand, such research needs to be related to current theoretical formulations of verbatim learning which emphasize the role of contextual stimuli (cf. Melton, 1967). On the other, there exists the possibility of further experimental exploration of different kinds of contexts, such as the "subsuming" and "correlative" types emphasized in Ausubel's (1967) theory. It seems likely that continuing to focus attention on the individual "fact" or "rule" will be a useful strategy in such research.

REFERENCES

AUSUBEL, D. P. The use of advance organizers in the learning and retention of meaningful verbal material. *Journal of Educational Psychology*, 1962, **53**, 243–249.

AUSUBEL, D. P. A cognitive-structure theory of school learning. In L. Siegel (Ed.), *Instruction: Some contemporary viewpoints*. San Francisco: Chandler, 1967.

BRIGGS, L. J., & REED, H. B. The curve of retention for substance material. *Journal of Experimental Psychology*, 1943, **32**, 513–517.

COFER, C. N. A comparison of logical and verbatim learning of prose passages of different lengths. *American Journal of Psychology*, 1941, **54**, 1–20.

DAVIDSON, R. E. Mediation and ability in paired-associate learning. *Journal of Educational Psychology*, 1964, **56**, 352–356.

DIETZE, A. G., & JONES, G. E. Factual memory of secondary school pupils for a short article which they read a single time. *Journal of Educational Psychology*, 1931, **22**, 586–598.

ENGLISH, H. B., WELBORN, E. L., & KILLIAN, C. D. Studies in substance memorization. *Journal of General Psychology*, 1934, **11**, 233–259.

GAGNÉ, R. M. *The conditions of learning*. New York: Holt, Rinehart & Winston, 1965.

HALL, J. F. Retroactive inhibition in meaningful material. *Journal of Educational Psychology*, 1955, **46**, 47–52.

JENSEN, A. R. *Individual differences in learning: Interference factor*. Berkeley, California: Institute of Human Learning, University of California, 1964. (Cooperative Research Project No. 1867).

KEPPEL, G. Verbal learning and memory. In, *Annual Review of Psychology*. Vol. 19. Palo Alto, Calif.: Annual Reviews, 1968.

KING, D. J., & RUSSELL, G. W. A comparison of

rote and meaningful learning of connected meaningful material. *Journal of Verbal Learning and Verbal Behavior,* 1966, **5,** 478–483.

McGOVERN, J. B. Extinction of associations in four transfer paradigms. *Psychological Monographs,* 1964, **78**(16, Whole No. 593).

MELTON, A. W. Individual differences and theoretical process variables: General comments on the conference. In R. M. Gagné (Ed.), *Learning and individual differences.* Columbus: Charles E. Merrill, 1967.

NEWMAN, E. B. Forgetting of meaningful material during sleep and waking. *American Journal of Psychology,* 1939, **52,** 65–71.

POSTMAN, L. Short-term memory and incidental learning. In A. W. Melton (Ed.), *Categories of human learning.* New York: Academic Press, 1964.

UNDERWOOD, B. J., & SCHULZ, R. W. *Meaningfulness and verbal learning.* Chicago: Lippincott, 1960.

YOAKAM, G. A. The effect of a single reading. In, *Report of the society's committee on silent reading, NSSE twentieth yearbook,* Part II. Chicago: National Society for the Study of Education, 1921.

(Received October 10, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 5, 362–369

EFFECTS OF ORAL AND ECHOIC RESPONSES IN BEGINNING READING[1]

MARY H. NEVILLE

Trinity and All Saints Colleges of Education, Leeds, England

96 Grade-1 children in 3 groups, at each of 2 learning aptitude levels, were compared to test the hypothesis: Giving an echoic or oral response before silent reading will, by encouraging the application of intonation patterns to beginners' reading, improve achievement and also reduce vocalization. 4 classroom teachers taught a preprimer vocabulary; *E* then taught the preprimer text to the groups, giving rigid silent, oral, or echoic training before silent reading. *S*s were tested for reading achievement and vocalization after each of the 3 preprimers. Analysis of variance indicated that echoic groups read more fluently and that echoic and oral training reduced vocalization. No significant group differences were found for word recognition and identification, or comprehension.

The view of reading as a code for the spoken language leads linguists to insist that the "first task of learning to read . . . is to teach the child to decode the written language to its language equivalents" (Levin, 1966, p. 140). By this they mean that all the language signals must be present in reading, including the paralinguistic intonation patterns. Since it has been hypothesized that intonation patterns help to show the connectedness and completeness of groups of words (Bolinger, 1957) and aid complex language learning (Braine, 1963), both Lloyd (1964) and Gliessman (1959) have suggested that intonation patterns, applied in reading, may also help the beginner unite singly identified words into units of thought. These hypotheses are supported by the relationship shown between the ability to apply normal intonation patterns to reading, and reading comprehension (Dearborn, Johnston, & Carmichael, 1949).

Linguists have given little guidance on how to teach the child the correspondences between written symbols and spoken language (Levin, 1966). Reading aloud should be helpful. It utilizes the secondary reinforcement properties of speech, makes possible extrinsic reinforcement, and, due to the learner's self-correction tendencies, enables him to bring his reading closer to the normal sounds of speech (Skinner, 1957). The association of meaning with the written words is also facilitated by an oral response (Hildreth, 1955; Osgood, Suci, & Tannenbaum, 1957) and, after practice, this meaning comes to be directly related to the written form. As pronunciation of words becomes superfluous, vocalization during silent reading should decrease and, by the process of cue reduction, silent reading will gradually develop (Anderson & Dearborn, 1952).

Through language learning, beginning readers have become accomplished in echoic behavior (Mowrer, 1960; Skinner, 1957) which they use to short-circuit the slower progressive approximation method of earlier language learning. Accurate echoing of speech has also acquired secondary reinforcing properties. The "look and say" reading method makes use of this accomplishment for teaching single words and could be extended to the teaching of larger language structures. If a model for reading were given, the child's oral, overt response would then become an echoic response which should be more efficient than the simple oral response for teaching the application of intonation patterns to reading.

The ability to organize singly identified words into a complete sentence should im-

[1] The data on which this paper is based were included in the author's dissertation, presented in partial fulfillment for the requirements of the PhD degree at the University of Calgary. The author is indebted to Ethel M. King, the supervisor for the study.

prove several aspects of reading. Fluent oral reading, defined (Harris, 1961) as reading where the words of the sentence are grouped in meaningful structural units indicated by appropriate pauses, stress, and inflections of voice, should occur. Reading comprehension should then improve and, with it, an increased use of context which should assist the recognition and identification of individual words (McCullough, 1958). With overt-response teaching methods, vocalization during silent reading should decline more quickly than when all responses to written words are silent.

In their study of overt and covert responses in reading, McNeil and Keislar (1963) suggest that the effects of an oral response may vary according to the learning aptitude of *S*s. Sex differences were also noted in an earlier study (Neville, 1965) for children who listened to a passage before reading it aloud.

The purpose of the present study was to investigate the effect of practicing an oral response before silent reading, or practicing an echoic response before silent reading, on the reading criteria of word recognition and identification, comprehension, fluency, and vocalization and to compare the effects of the oral and echoic responses at two learning-aptitude levels and for each sex. The hypothesis was: Practicing an echoic response before silent reading would give the best performance on all criteria while practicing an oral response before silent reading would give a better performance than no overt practice before silent reading.

METHOD

The experiment was performed in one large school with four complete Grade-1 classes in a lower-middle/upper-lower class area of Calgary.[2] Although all 110 Grade-1 pupils were taught in the experimental classes, 6 of them were not considered as *S*s because they were either repeating the grade, could not speak English, or were outside the chronological-age range of from 5 years 8 months to 6 years 8 months. No child had learned to read before entering school.

During the first 3 weeks of school, three tests predictive of reading success[3] were administered. As the predictive values for the three tests are very similar (Olson, 1958) and the score distributions were close to normal, scores were converted to z scores and added together to form, for each *S*, a composite score. Two groups of equal size were formed by assigning *S*s above and below the composite score median to upper and lower learning-aptitude levels.

For each level, *S*s were randomly assigned to three treatment groups per sex. The *S*s in each of these six groups at each learning level were then randomly assigned to two classes. Thus, at each learning level there were two classes, each with three equal-sized treatment groups. The three treatment groups, combined over the two classes of a level, were comparable with regard to sex distribution. The teachers most suited to teach a particular level were chosen by the assistant principal: At each level the two teachers were randomly assigned to classes. All four teachers had had 2 or more years' experience in teaching Grade 1.

The normal reading texts (*Reading for Meaning Readers*, McKee, et al., 1957) and program used by the Calgary Public School Board were followed in the study. All *S*s had worked through the same prereading program and continued then, in the fifth week of school, into the associated preprimer reading material. In the upper-learning-level classes, the classroom teachers taught the entire vocabulary of the first preprimer at an average rate of 2 words a day, this rate being increased to 2.5 words a day for the two subsequent preprimers. The two lower-learning-level classes took one-third as much time again to learn the same words for each preprimer. To present the words the teachers used the usual "look and say" method outlined in detail in the teacher's manual but the amount of context used in presenting each word was reduced to one very short sentence. At no time in the classroom was any other connected material read by *S*s.

After the preprimer vocabulary had been taught by the teacher, *E* taught the reading of each preprimer for 10 consecutive school days, keeping the time constant over levels. Reading was taught both morning and afternoon for all *S*s, 3 pages of the 60-page preprimer being presented per session.

At each learning level, *S*s of the same treatment group were combined over the two classes and taught, or trained, together in a room near the Grade-1 classrooms. There were 17 or 18 *S*s in each group plus 1 "non-*S*." While a group was absent from the classroom the teacher continued with

[2] The author would like to thank R. Warren, Superintendent, Calgary Public Schools, the school principal, M. Gutiw, and the assistant principal, Winfred Hargreaves for their cooperation during this research. She is especially grateful to the four Grade-1 teachers, Sibyl Faid, Marian Hansen, Marg. Peters, and Christina Roberts.

[3] Pintner-Cunningham Primary General Ability Test: Verbal Series, Form A; Harrison-Stroud Reading Readiness Profiles, Auditory Subtests 4 and 5; Murphy-Durrell Reading Readiness Analysis, learning rate subtest.

work other than the presentation of preprimer vocabulary.

Training

Three methods followed by E as she taught the reading of the three preprimers constituted the treatments: silent (no oral response before silent reading), oral (oral response before silent reading), and echoic (echoic response before silent reading). The silent training approximated the method of the preprimer guidebook, and so the silent group could be considered as a control group. Training for any treatment did not vary between pages, preprimers, or levels. Each group followed the same treatment for each preprimer.

The invariable procedure for each page of reading (outlined in Table 1) was as follows: Each page was introduced for all groups by picture reading. The children's books were turned face down on their knees and only the picture in E's book, with the words masked, was discussed. The E always asked, "Who is in the big picture? What is ——— doing?" Books were then turned up and reading began.

Before the first presentation of the page to the silent group and the third presentations to the oral and echoic groups (all silent reading—Table 1), E asked the same comprehension question to guide the reading. She then added, "Read with your eyes and do not move your lips."

For the second, oral presentation for the oral and echoic groups, E said, "Read out loud and make your reading sound just like quiet talking." For the second, silent presentation for the silent group they were told, "Read it again, silently. Do not move your lips. While you read, *think* how it would sound if it were just like talking."

For the first and third presentations, respectively, the echoic and silent groups were told, "Look at the first word on the page. Now listen carefully to me (or child's name in the silent group) and watch the words in the book. (Story character) is talking." The E said for the first, oral presentation of the oral group, "Read the story out loud. (Story character) is talking."

The Ss remained seated in front of E during all oral reading. They read with normal speech volume but at differing rates so that reading was not in unison. No distraction of any one child by the voices of the others was observed.

TABLE 1

TREATMENT METHODS USED TO TEACH 1 PAGE OF READING

Presentation of text	Method		
	Silent	Oral	Echoic
First	Silent reading	Oral response	Listen to model read by E
Second	Silent reading	Oral response	Echoic response
Third	Listen to oral reading by one S of group	Silent reading	Silent reading

In all groups, if an S did not know a word, E told it to him. Words giving general difficulty were told to all groups and written on the blackboard.

Since all the words were taught before Ss had their experimental training, if an S had been absent not more than 5 consecutive school days he continued on his return with his treatment group and the missed practice was made up in individual sessions. Such a procedure was unavoidable since two-thirds of the Ss were absent at some time during the 60 group-training sessions.

Testing

At the end of each preprimer every S was tested individually by an E-constructed test passage pertaining to the preprimer. Each of the three tests consisted of a story using all the words of the particular preprimer, as well as other words from earlier preprimer vocabularies plus several new words. The only picture was of the preprimer character who was "talking" in the story. Ten associated comprehension questions were also constructed for each story, five to test simple recall and five to test inferred understanding. The test passage, typed on quarto paper with a primary typewriter, was placed on a specially constructed reading stand on a child-sized table. Attached to the front of the stand was a rigid section on which S rested his chin. This section contained a very sensitive microphone, concealed just below a net-covered aperture, and connected to a Uher 4000 Report-L tape recorder which recorded voice sounds. No child indicated that he suspected the presence of a microphone.

Each S, taken in random order, sat in front of the stand with E beside him. She explained that S was to read a story and that after it had been read, questions would be asked about it. The S was asked to read silently, and told to read with his eyes and not move his lips. No words were identified for S during this reading. The comprehension questions were then asked orally and the answers recorded verbatim.

The E next asked S to read the story again, this time out loud so that his reading would sound "just like talking." As S read orally, words that he did not recognize or identify were told to him after a pause of approximately 5 seconds and his errors recorded after the scoring method of the Gray Oral Reading Test (Gray, 1963). This testing procedure was followed for each of the three preprimer tests.

The Grade-1 teachers marked the comprehension questions using model answers prepared by E on the basis of a trial of the material in the previous school year. The names on the record sheets were concealed. Three "number correct" scores were obtained for each test: literal comprehension, inference comprehension, and total comprehension scores.

When any speech sounds were discerned by E

on the silent reading recordings, the tapes were listened to by *E* and two other raters. Vocalized words were totaled over the test passage and averaged over the three raters to obtain a vocalization score. Interscorer reliability based on the average intercorrelation among the three sets of ratings was .48.

The *E* rated all the oral reading recordings for fluency, as defined in the introduction, and obtained a score of the total number of nonfluent sentences of more than one word. Because *Ss* had had such limited reading experience, pauses for word identification, as well as word substitutions or additions, were ignored when rating. As a check on the reliability of *E*'s rating, two specialists in primary reading also listened to the reading of 20 *Ss* chosen randomly from all groups, for each of the three test passages. Interrater reliability for nonfluent sentences over the 60 passages, based on the average intercorrelation among the three raters, was .45.

The word-recognition error score was the number of words marked in each error category; the word-identification score was the number of new words read correctly in the passage.

At the end of the third preprimer testing, the teachers resumed the normal classroom teaching of reading. Most children were ready to start the first primer of the reader series.

At each level, 8 weeks after the end of the experiment, *Ss* were posttested with the Primary Reading Profiles, Level I, a group reading-achievement test edited by the senior author of the textbooks used in the experiment. Progress in comprehension, word identification from context, and word recognition were measured; fluency and vocalization were not retested.

RESULTS

A three-factor, Lindquist Type III (Linquist, 1953) analysis of variance was used for the main analyses. The between factors for the analysis of each dependent variable were treatment groups and learning-aptitude levels. The three preprimer tests constituted the within factor. In a second analysis of each dependent variable the two between factors were treatment groups and sex differences, the within factor, as before, being the three tests.

The posttest data for word recognition, word identification, and comprehension were analyzed separately by two-factor analyses of treatment group by learning-aptitude level, and treatment group by sex difference.

Some *Ss* were lost during the experiment and others were randomly rejected to give 16 in each treatment group at each level for preprimer and posttest data. For the analysis by sex difference, three more *Ss* were rejected so that each treatment group contained 17 girls and 14 boys.

Comparison of Treatments by Learning Levels

Means and standard deviations for all the preprimer test variables, averaged over the three preprimer tests, are given in Table 2. Table 3 shows the levels of significance for the Type III analyses of variance for each of these variables.

Tests of homogeneity of variance showed a marked heterogeneity of variance for vocalization only (between treatment groups). Accordingly, following the recommendation of Norton's study (Lindquist, 1953, p. 86), the required level of significance for this one test was raised to .025 (as opposed to .05 for all other tests) and a *t* test for unequal variances was used.

The significant main effect for performance over the three preprimer tests was shown by *t* tests to be in the expected direction for word-recognition errors and nonfluent sentences. The significant effect for all comprehension was due to a decline in performance over tests: For word identification, the effect was caused by higher performance on the second preprimer test in comparison with the other two.

The learning-levels effect for word identification, literal comprehension, and fluency was in the expected direction.

The significant treatments effect for vocalization was shown by *t* tests to be due to superior performance of the oral group and echoic group compared to the silent group. The significant Tests × Treatments interaction for nonfluent sentences was caused by a significantly better performance of the echoic group in relation to the silent group on the third preprimer test.

Thus the first hypothesis of superiority of the oral response practice over no overt practice was accepted only for vocalization. The hypothesis of superiority of echoic practice over both oral response practice and no overt practice was not accepted for any dependent variable, but for

TABLE 2

MEANS AND STANDARD DEVIATIONS FOR THREE TREATMENT GROUPS AT TWO LEARNING LEVELS FOR ALL PREPRIMER TESTS

Test	Learning level	Treatment					
		Silent		Oral		Echoic	
		M	*SD*	*M*	*SD*	*M*	*SD*
Word recognition(errors)	Upper	17.75	14.28	14.85	15.79	13.48	10.85
	Lower	13.96	17.79	16.00	11.52	17.06	12.55
Word identification	Upper	1.37	1.19	1.70	1.27	1.59	1.02
	Lower	1.00	1.06	0.77	0.66	1.44	1.20
Literal comprehension	Upper	2.40	0.44	2.42	0.64	2.58	0.87
	Lower	2.10	0.70	1.90	0.42	2.33	0.42
Inference comprehension	Upper	1.85	0.82	1.69	0.66	1.67	0.63
	Lower	1.52	0.73	1.63	0.69	1.79	0.63
Total comprehension	Upper	4.25	1.06	4.10	1.04	4.25	1.31
	Lower	3.63	1.21	3.52	0.82	4.13	0.93
Nonfluent sentences	Upper	6.00	1.15	5.08	1.92	5.44	1.58
	Lower	6.46	1.51	6.42	1.49	5.83	1.51
Vocalization	Upper	15.52	18.25	2.17	4.03	7.90	11.25
	Lower	12.46	15.35	3.83	6.83	5.08	9.60

Note.—$N = 16$ at each learning level.

both vocalization and fluency the echoic group was superior to the silent group.

Comparison of Treatments by Sex

There was no main sex effect for any of the preprimer tests but interaction between treatments and sex differences was found for all comprehension tests and for word recognition. It was found from t tests that, for word recognition, the boys of the oral group were superior to all other groups of either sex for the first preprimer test.

TABLE 3

TYPE III ANALYSIS OF VARIANCE FOR ALL DEPENDENT VARIABLES

Source of variance	*df*	Dependent variables													
		Word recognition		Word identification		Literal compre-hension		Inference compre-hension		Total compre-hension		Fluency		Vocali-zation	
		MS	*F*	*MS*	*F*	*MS*	*F*	*MS*	*F*	*MS*	*F*	*MS*	*F*	*MS*	*F*
Between-*S*	95	602.39		380.99		1.28		1.51		3.77		7.97		498.46	
Group (B)	2	8.78	0.01	302.68	0.82	2.29	1.94	0.13	0.08	3.50	0.95	9.52	1.25	3027.17	6.68**
Level (C)	1	7.06	0.01	1716.00	4.66*	9.03	7.64**	0.59	0.38	14.22	3.84	38.28	5.02*	141.68	0.31
B × C	2	338.81	0.54	382.60	1.04	0.51	0.43	1.27	0.82	1.85	0.50	6.59	0.86	169.96	0.38
Error	90	628.05		367.87		1.18		1.56		3.70		7.63		453.53	
Within-*S*	192	67.42		216.74		1.80		1.05		3.57		3.66		222.75	
Tests (A)	2	1560.47	28.98***	1878.58	9.21***	82.13	84.54***	4.85	4.62*	122.64	51.77***	186.98	111.68***	128.76	0.59
A × B	4	9.08	0.17	49.27	0.24	0.85	0.87	0.10	0.09	0.73	0.31	5.46	3.26*	106.05	0.49
A × C	2	9.66	0.18	1.53	0.01	1.64	1.68	1.26	1.21	5.42	2.29	1.67	1.13	886.68	4.08*
A × B × C	4	18.84	0.35	233.82	1.15	0.52	0.05	0.14	0.13	0.14	0.06	0.43	0.26	308.95	1.42
Error	180	53.85		204.01		0.97		1.05		2.37		1.67		217.09	
Total	287														

* $p < .05$.
** $p < .01$.
*** $p < .001$.

With all the comprehension tests, interaction was symmetric. Girls of the echoic group scored high, while boys of this group scored low; for the silent group, boys scored high and girls low; in the oral group, performance was similar for boys and girls. This effect, except for inference comprehension, was due largely to the results of the third preprimer test.

Reading Posttest

There was a levels effect in the expected direction for word recognition and identification but not for comprehension. There was no treatments effect for any test, nor interaction between levels and treatments. Neither were any significant sex differences found.

DISCUSSION

Practice in giving an echoic or oral response before reading silently caused no improvement in word recognition, word identification, or comprehension. This is not in agreement with the earlier study (Neville, 1966) where echoic *S*s were superior, or with the experiment of McNeil and Keislar (1963) using individual reading practice booths, where an oral response group was superior in silent reading to a covert response group. But neither did the rather passive listening period prior to the echoic response seem to have the adverse effect on reading postulated by Buswell (1922) or Duggins (1958).

One possible reason for the lack of significant differences may have been the nature of the silent reading. In the training periods, true silent reading by the silent group was very difficult to achieve as the children whispered whenever they thought they were unobserved. Hildreth (1955) believes that the beginner virtually cannot read silently, so strong is the connection between printed words and speech: In the present study, the social interaction (Mace, 1966) of the group reading situation may also have encouraged imitation of vocalization.

The actual vocalization scores of the echoic group appear, on inspection of Table 2, to lie between those of the oral and silent groups but, because of the unequal variances of the groups, the only significant differences were between the oral and silent, and echoic and silent groups. Treatments effect showed at the first test and thereafter changed little. During the training periods, too, the lack of vocalization in the echoic and oral groups contrasted sharply with the vocalizing of the silent group. It is possible that the extra, legitimate practice in associating the sounds of words with their symbols hastened the process of cue reduction and reduced vocalization in the oral and echoic groups.

Although giving an oral response improved fluency of oral reading at the start of the experiment, this effect did not continue. During the third preprimer, with some of the better oral-group readers, a monotone chanting developed, perhaps as a reaction against the boredom of so much oral reading. The echoic group showed continuing improvement which, by the third test, gave scores significantly superior to the silent group.

Increased fluency did not produce the predicted improvement in other reading skills, particularly comprehension, but possibly the effect would have shown with a still longer experiment. Alternatively, the difficulty of the comprehension questions and a tendency to guess their answers may have obscured treatment differences: The decline in comprehension scores over the three tests appeared to be due to too rapid an increase in difficulty.

Contrary to McNeil and Keislar's (1963) suggestion, learning level did not affect the success of the oral response group—or of any other treatment group. The lack of simple levels effect in the word recognition results was to be expected since all the children supposedly knew the words of the preprimers before coming for training. In the posttest comprehension and the inference comprehension, the levels effect was possibly obscured by the low scores of all pupils on these, the most difficult tests.

In discussing sex differences which, in beginning reading, usually favor girls, McNeil (1964) suggests that they may occur because boys receive more negative comments than girls during reading and fewer opportunities to read. Possibly the very

rigid form of all reading lessons in the experiment reduced these effects. The interaction with word recognition, showing that boys from the oral group performed better initially, should be viewed very cautiously and requires further study. The sex interaction for comprehension is interesting in that just the opposite effect was found in the pilot study (Neville, 1965). The boys' success in the earlier study was postulated to be due to their superior listening ability. Possibly the boys of the echoic group in the present study continued to rely on their good listening comprehension when answering the comprehension questions, while the girls paid more attention to the reading. Silent reading practice, which necessitated attention to the text for full comprehension, was perhaps better training for the boys for reading comprehension, especially of the more difficult third test passage.

Some features of the experimental design may be of interest. The study attempted to carry out research in a situation approximating the classroom, at the same time trying to control some of the variables thought to confound much classroom research. (*a*) To keep the teacher's influence the same for all groups during training, *E* taught the whole sample. There were still teacher variables operating, especially in the classroom teaching of words (although no significant differences were found between the word-recognition scores of classes of the same level) but these effects were equalized within each treatment group. (*b*) The Hawthorne effect was comparable for all groups including the silent or control group. (*c*) The reading methods were prescribed precisely and rigidly followed during training, in the hope that treatments effect could be fairly attributed to the one differing variable, type of response practice. The children did not object to the unvaried form of the training and at the end of the experiment only six children, all good readers from all three groups, signified that they disliked reading. (*d*) The research continued training and testing over a reasonable length of time. The chief problem revealed by the study is the difficulty of gaining a reliable assessment of such aspects of beginning reading as comprehension and fluency when *S*s have so rudimentary a reading ability.

The reduced vocalization and improved fluency of the overt response groups give some support to the linguists' contention that the child needs to convert written words to familiar speech and suggest that the reading program should include oral and echoic responses as well as silent reading. Moreover, the natural tendency for beginners to vocalize should not cause concern since, at this stage, overt responses appear to foster reading achievement.

REFERENCES

ANDERSON, I. H., & DEARBORN, W. F. *The psychology of teaching reading.* New York: Ronald Press, 1952.

BOLINGER, D. L. Intonation and grammar. *Language Learning*, 1957, **13**, 31–38.

BRAINE, M. D. On learning the grammatical order of words. *Psychological Review*, 1963, **70**, 323–348.

BUSWELL, G. T. Fundamental reading habits: A study of their development. *Supplementary Educational Monographs*, 1922, No. 21.

DEARBORN, W. F., JOHNSTON, P. W., & CARMICHAEL, L. Oral stress and meaning in printed materials. *Science*, 1949, **110**, 404.

DUGGINS, L. A. Theory and techniques of auditory perception as an approach to reading. In O. S. Causey (Ed.), *The reading teacher's reader.* New York: Ronald Press, 1958.

GLIESSMAN, D. Understanding reading from the viewpoint of sentence psychology. *Reading Teacher*, 1959–60, **13**, 22–28.

GRAY, W. S. *Gray Oral Reading Tests*, (Ed. by H. M. Robinson) New York: Bobbs-Merrill, 1963.

HARRIS, A. J. *How to increase reading ability.* (4th ed.) New York: Holt, Rinehart & Winston, 1961.

HILDRETH, G. The role of pronouncing and sounding in learning to read. *Elementary School Journal*, 1955, **55**, 141–147.

LEVIN, H. Reading research: What, why, and for whom? *Elementary English*, 1966, **43**, 138–147.

LINDQUIST, E. F. *Design and analysis of experiments in psychology and education.* Boston: Houghton Mifflin, 1953.

LLOYD, D. J. Intonation and reading. *Education*, 1964, **84**, 538–541.

MCCULLOUGH, C. M. Context aids in reading. *Reading Teacher*, 1958, **11**, 225–230.

MACE, L. L. Sequence of vocal response-differentiation training and auditory stimulus-discrimination training in beginning French. *Journal of Educational Psychology*, 1966, **57**, 102–108.

McKEE, P. M., ET AL. *Reading for meaning readers.* Boston: Houghton Mifflin, 1957.

McNEIL, J. D. Programmed instruction versus usual classroom procedures in teaching boys to read. *American Educational Research Journal,* 1964, **1,** 113–119.

McNEIL, J. D., & KEISLAR, E. R. Value of the oral response in beginning reading: An experimental study using programmed instruction. *British Journal of Educational Psychology,* 1963, **33,** 162–168.

MOWRER, O. H. *Learning theory and the symbolic processes.* New York: Wiley, 1960.

NEVILLE, M. H. The effect of silent reading, oral reading, and listening on accuracy and comprehension in beginning reading. Unpublished master's thesis, University of Calgary, 1965.

NEVILLE, M. H. Methods of teaching reading to beginners. *Alberta Journal of Educational Research,* 1966, **12,** 131–139.

OLSON, A. V. Growth in word perception abilities as it relates to success in beginning reading. *Journal of Education,* 1958, **140,** 25–36.

OSGOOD, C. E., SUCI, G. J., & TANNENBAUM, P. H. *The measurement of meaning.* Urbana: University of Illinois Press, 1957.

SKINNER, B. F. *Verbal behavior.* New York: Appleton-Century-Crofts, 1957.

(Received October 24, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 5, 370–374

INTERACTION EFFECTS OF ITEM-DIFFICULTY SEQUENCE AND ACHIEVEMENT-ANXIETY REACTION ON ACADEMIC PERFORMANCE

DAVID C. MUNZ AND ALBERT D. SMOUSE
University of Oklahoma

In order to test the assumption that individual reaction to test taking mediates the effect of item-difficulty sequence on performance, college freshmen were randomly assigned a final examination with items sequenced either hard to easy (H-E), easy to hard (E-H), or at random (R), and were then classified within each sequence as to achievement-anxiety type. As predicted, a 3 (item-sequence) × 4 (reaction-type) analysis of variance ($N = 120$) yielded significant F ratios ($p < .01$) only for reaction type and interaction; however, several specific performances were significant in the nonpredicted direction. Results are explained using the inverted-U hypothesis and the assumption that item sequences are progressively more arousing in the order of R, E-H, H-E. Implications of the research are discussed.

Recent investigations of the standard test construction practice of arranging test items in an order of increasing difficulty, that is, the easier items first followed by the progressively more difficult ones, have found no empirical justification for such a practice (Brenner, 1964; Smouse & Munz, 1968). There appears to be little influence of an easy-to-hard item-difficulty sequence, as compared with a hard-to-easy or random sequence, on academic achievement scores when group measures are used. However, no attempt has been made to investigate the effect on performance scores of an interaction of personality factors typically found in the test-taking situation and different item-difficulty arrangements of test items.

There is an abundance of literature on personality factors which influence test-taking behavior in the classroom. One such factor is the differential reactions of individuals to test-taking anxiety. Until recently many investigators viewed anxiety as a unidimensional personality trait. Alpert and Haber, authors of the Achievement Anxiety Test (AAT; 1960), have preferred to view test-taking anxiety (achievement anxiey) as a bidimensional construct which may have facilitating as well as debilitating effects on academic performance. For some individuals an anxiety-provoking situation, such as a typical college examination, facilitates their performance while for others it depresses performance. Further, there are those individuals whose test performance is not affected by anxiety-provoking situations, either by improving or depressing their scores. "Thus, an individual may possess a large amount of both anxieties or of one but not the other, or of none of either [p. 213]."

The major argument supporting the arrangement of test items in an easy-to-hard (E-H) difficulty sequence has been that an E-H arrangement decreases test-taking anxiety, thereby facilitating performance. In comparison, a random (R) arrangement does not affect test-taking anxiety, and therefore does not influence test performance, while a hard-to-easy (H-E) arrangement increases test-taking anxiety thereby depressing performance. The purpose of this study was to investigate the notion that item-difficulty arrangement does significantly affect performance scores but only by interacting with test-taking personality factors. More specifically, differential reactions to test-taking anxiety, as measured by the AAT, interacting with three item-difficulty arrangements of test items (E-H, H-E, and R) has an effect on achievement-test scores. It was hypothesized that because of the anxiety-generating effects of the H-E arrangement of test items those individuals whose performance is improved under anxiety-provoking situations score significantly higher on the H-E sequence than those individuals whose

performance is depressed or not affected by test anxiety. Further, because of the anxiety-reducing effects of an E-H arrangement those individuals whose performance is depressed under anxiety-provoking situations score significantly higher on the E-H sequence then those individuals whose performance is facilitated or not affected by test anxiety. Moreover, individuals whose performance is not affected by test anxiety are also not affected by item-difficulty arrangement. The research reported here was designed to test the following specific hypotheses:

Hypothesis 1. Item-difficulty arrangement of test items, as a main effect, does not significantly affect performance scores. This hypothesis is in accordance with recent findings.

Hypothesis 2. Differential reactions to test-taking anxiety, as measured by the AAT, significantly affect performance scores.

Hypothesis 3. Item-difficulty sequence (E-H, H-E, and R) and achievement-anxiety reaction types (facilitators, debilitators, nonaffecteds, and high-affecteds) interact to produce a significant effect on performance scores.

Hypothesis 3a. Facilitators, those *Ss* scoring relatively high on the facilitating scale (AAT+) and relatively low on the debilitation scale (AAT−), perform significantly better on the H-E arrangement than do the other three reaction types.

Hypothesis 3b. Debilitators, those *Ss* scoring relatively high on AAT− and relatively low on AAT+, perform significantly better on the E-H arrangement than do the remaining three reaction types.

Hypothesis 3c. Nonaffecteds', Ss low on both AAT+ and AAT−, performance is not significantly affected by item-difficulty arrangement. No further predictions were made regarding high-affecteds, those *Ss* scoring high on both AAT+ and AAT−.

METHOD

Subjects

The *Ss* were 120 male and female students enrolled in four sections of an introductory psychology course taught at the University of Oklahoma. The *Ss* were chosen from 181 students who had filled out the AAT which had been presented as a research project of the psychology department. The *Ss* were informed that the information procured from the questionnaire would be used for research purposes only.

Instruments

The AAT was designed to measure the reported effects of anxiety experienced in test-taking situations. This instrument distinguishes between different degrees of anxiety that is reported by the respondent as either facilitating or debilitating to test performance. Each type of anxiety is measured by a separate subtest of items (AAT+ scale and AAT− scale) together comprising a 19-item questionnaire.

The second independent variable consisted of three sequences of items on a final examination for an introductory psychology course. These sequences, constructed for use in another study (Smouse & Munz, 1968), contained the same 100 multiple-choice items (four alternatives) differing only in item-difficulty order (H-E, E-H, and R). The 100 items were selected from a pool of 197 items administered as a final examination to 931 students in the previous semester's introductory psychology course. Item analysis of the 197 items ($N = 931$) yielded item-difficulty values defined as percentage of students passing a given item (Nunnally, 1959). Those 100 items selected from the pool and having a relatively even spread of item-difficulty values over a range of 17.7–96.6% were arranged in the three item-difficulty orders: E-H, H-E, and R.

Procedure

The AAT was administered in class to four sections of an introductory psychology course 1 week prior to the final examination. Three forms of the final examination, H-E, E-H, and R, were administered to all sections of the introductory psychology course and were randomly distributed within each section. Each section took the final exam under its own instructor and was given as much time as needed to complete the examination. The students recorded their answers on answer sheets which were electronically scored.

The achievement-anxiety types were constructed by selecting *Ss* from each item-arrangement group in the following manner. An AAT+ score and an AAT− were obtained on each *S* after which the AAT− score was subtracted from the AAT+. A positive difference indicated a relatively high AAT+ and a negative difference indicated a relatively high AAT−. When these difference scores were ranked, the top 10 *Ss* in the distribution were defined as facilitators and the bottom scoring *Ss* were defined as debilitators. For all remaining *Ss* the two scores were summed and ranked. The top 10 *Ss* in the resulting distribution were defined as high-affecteds while the bottom 10 scores were defined as nonaffecteds.

The final examination data (total number of items answered correctly) for the 120 *Ss* were sub-

jected to a 3 (item-difficulty order) × 4 (achievement-anxiety types) analysis of variance.

RESULTS

Table 1 presents the analysis of variance results for the total sample along with the simple main effects analysis. Consistent with Hypothesis 1, the analysis revealed no statistically significant item-difficulty order effect upon performance scores. There was, however, a statistically significant effect of the achievement-anxiety reaction variable on performance scores ($F = 4.32$, $p < .01$). These results supported Hypothesis 2. Probing with the Neuman-Kuels Test (NKT) revealed that only the difference between the facilitators and the debilitators ($p < .01$) and the difference between the facilitators and nonaffecteds ($p < .01$) were significant, facilitators scoring higher in both instances.

There was a statistically significant interaction among the three item-difficulty orders and the four achievement-anxiety types ($F = 3.22$, $p < .01$); hence, Hypothesis 3 was supported. A simple main effects analysis indicated that within the achievement-anxiety factor the R arrangement and E-H arrangements were significant ($F = 5.36$, $p < .01$; $F = 4$, $p < .01$, respectively). Examining these two item-difficulty orders with the NKT revealed that (*a*) on the R form, facilitators and high-affecteds scored significantly higher than the debilitators ($p < .05$) and nonaffecteds, ($p < .01$), and (*b*) on the E-H form, facilitators scored significantly higher than the other three anxiety types ($p < .05$; see Figure 1). The specific interaction findings contradict the subhypotheses formulated under Hypothesis 3.

TABLE 1

ANALYSIS OF VARIANCE SUMMARY FOR PERFORMANCE SCORES AS A FUNCTION OF ACHIEVEMENT-ANXIETY TYPE AND ITEM-DIFFICULTY SEQUENCE

Source of variation	*df*	*MS*	*F*
Item-difficulty sequences (A)	2	107.50	1.05
Achievement-anxiety types (B)	3	442.41	4.32*
B for A_1	3	582.50	5.36*
B for A_2	3	409.37	4.00*
B for A_3	3	108.63	1.06
A × B	6	329.04	3.22*
Error	108	102.32	
Total	119		

Note.—Abbreviated: A_1 = random sequence, A_2 = easy-hard sequence, A_3 = hard-easy sequence.

* $p < .01$.

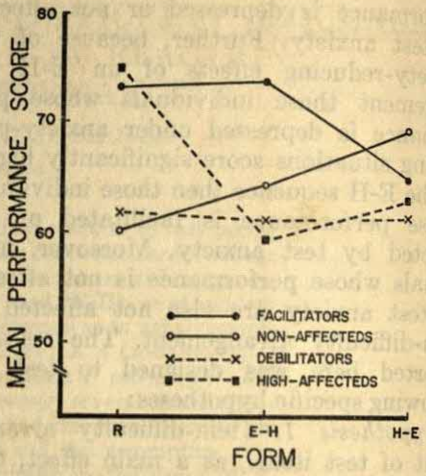


FIG. 1. Mean performance scores for achievement-anxiety types (facilitators, debilitators, high-affecteds, and nonaffecteds) on item-difficulty arrangements consisting of random (R), easy to hard (E-H), and hard to easy (H-E).

DISCUSSION

In accordance with recent investigations, the results of this study suggest that the standard test-construction practice of arranging test items in an order of increasing difficulty is not justified from the argument that an E-H item-difficulty sequence produces higher performance scores than an R or H-E arrangement. There appears to exist a more complicated relationship between item-difficulty arrangement of test items, specific personality correlates, and achievement-test performance.

Several of the interaction effects were unexpected. First, there were no significant differences among reaction types on the H-E form. Hypothesis 3a had predicted facilitators' performance to be superior to those of the other reaction types. Second, for the E-H form, the only significantly superior performance was that of the facilitators, not the debilitators as had been predicted in Hypothesis 3b. Third, although

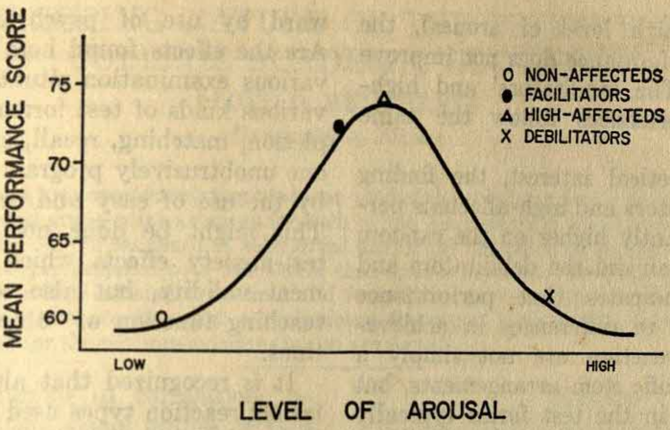


FIG. 2. Curve representing the inverted-U hypothesis, here showing achievement test performance as a function of item-difficulty sequence. Plottings superimposed on the theoretical curve show mean performances of the various anxiety-reaction types for the random sequence.

no predictions were made with regard to the ranking of mean performances within the R sequence, it was nevertheless surprising to find the facilitators and high-affecteds clustering significantly above the debilitators and nonaffecteds. Although, strictly speaking, the interaction hypothesis (Hypothesis 3) was supported inasmuch as the mean performances of the various reaction types fluctuated differentially from sequence to sequence, the pattern of fluctuations clearly calls for an alternate set of hypotheses. One plausible post hoc explanation can be made on the basis of two assumptions. The first involves the inverted-U function, and the second has to do with the relative arousal potentials of the three item-difficulty forms.

The inverted-U hypothesis states that behavioral efficiency varies as a curvilinear function of what has been variously referred to as "arousal" (Malmo, 1959), "drive level" (Easterbrook, 1959), and "activation level" (Fiske & Maddi, 1961). This function, shaped roughly like an inverted U, implies that there is a degree of arousal which is optimal for performing a given task. If an individual or group of individuals are functioning at a drive level which is greater or less than optimum for a particular task, then performance on that task is impaired. Thus, if one assumes that the differential reactions to achievement-test anxiety on the R form places the reaction types on the performance curve as shown in Figure 2, this assumption alone will explain why the facilitators performed significantly higher on the examination than the debilitators and nonaffecteds, and why the nonaffecteds performed at the same level as the debilitators. The significantly superior performance of the high-affecteds over the nonaffecteds and debilitators is consistent with the placement of the high-affecteds higher on the performance curve. The inverted-U hypothesis becomes even more plausible when called upon to explain the total interaction effects.

The unexpected interaction findings can be explained if the second assumption is made, namely, that each of the item arrangements produces a different degree of arousal, increasing in the order of R, E-H, and H-E. If one combines this assumption with the assumption that the various achievement-anxiety reaction types lie initially on the performance curve as shown in Figure 2, then as one moves in the direction of increasing stress, that is, from R to E-H to H-E, 11 of the 12 results plotted in Figure 1 are predictable (compare Figures 1 and 2). Since the nonaffecteds' performance is lower than the facilitators' and high-affecteds' performance due to their low level of arousal, their performance improves proceeding across test forms. Since the debilitators' performance is lower than the facilitators' and high-affecteds' performances due to the debili-

tators overly high level of arousal, the debilitators' performance does not improve across forms. The facilitators' and high-affecteds' performances follow the same rationale.

Of more practical interest, the finding that the facilitators and high-affecteds performed significantly higher on the random arrangement than did the debilitators and nonaffecteds indicates that performance differences due to differences in achievement-anxiety reaction are not simply a product of specific item arrangements, but probably exist in the test forms typically found in the classroom. Further, the E-H order favors primarily the nonaffecteds while actually depressing the high-affecteds' performance. It is recognized, however, that the arousal value of the E-H form may lie in the clustering of difficult items at the end of the exam so that the practice of placing a few easy items at the beginning of an otherwise randomly ordered test may not have the same dramatic effect. But the implications of R sequencing for classroom grading must be contended with. If this study is generally valid, then the H-E sequence provides least variance attributable to personality factors and should be used when one is attempting to assess only academic achievement. If the personality variables in question become a legitimate part of the assessment, then the sequence should be selected accordingly.

In addition to serving as a reminder of the weaknesses inherent in the intuitive approach to test construction, results of this study suggest several subsequent avenues of research. First, how does one manipulate the debilitators' performance upward by use of psychometric strategy? Are the effects found here constant across various examination situations and across various kinds of test formats such as completion, matching, recall, etc.? Also, might one unobtrusively program an examination by the use of easy and/or difficult items? This might be done not only to reduce test-anxiety effects which lower achievement validity, but also to maximize the teaching function of "objective" examinations.

It is recognized that although the test-taking reaction types used as an independent variable in this study are based on a theoretical instrument, they have been somewhat operationalized and hence need theoretical support. Further research would be needed, however, to establish such theoretical support on an ad hoc basis, and this is presently being pursued.

REFERENCES

ALPERT, R., & HABER, R. N. Anxiety in academic achievement situations. *Journal of Abnormal and Social Psychology*, 1960, **61**, 207–215.

BRENNER, M. H. Test difficulty, reliability, and discrimination as functions of item difficulty order. *Journal of Applied Psychology*, 1964, **48**, 98–100.

EASTERBROOK, J. A. The effect of emotion on cue utilization and the organization of behavior. *Psychological Review*, 1959, **66**, 186–201.

FISKE, D. W., & MADDI, S. R. (Eds.) *Functions of varied experience*. Homewood, Ill.: Dorsey Press, 1961.

MALMO, R. B. Activation: A neurophysiological dimension. *Psychological Review*, 1959, **66**, 367–386.

NUNNALLY, J. C. *Tests and measurements*. New York: McGraw-Hill, 1959.

SMOUSE, A. D., & MUNZ, D. C. The effects of anxiety and item difficulty sequence on achievement testing scores. *Journal of Psychology*, 1968, **68**, 181–184.

(Received October 25, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 5, 375–380

THE FLEETING NATURE OF THE PREDICTION OF COLLEGE ACADEMIC SUCCESS[1]

LLOYD G. HUMPHREYS
University of Illinois

Data are reported relating freshmen predictors to independently computed grade-point averages for each of the 8 semesters of undergraduate college residence. A substantial amount of instability of intellectual performance over this 4-year time span is revealed. Implications for college admission research and for policies governing failure and probation are discussed. A preliminary theoretical basis for the data and for the recommendations is also developed.

There are hundreds of correlations reported between high school grades or rank in class, college entrance examinations, and the criterion of college grades. Many of these studies use first quarter or first semester averages, a good many others use first year averages, and a few use 4-year cumulative averages. Studies using freshmen predictors and senior grades are almost nonexistent. Several years ago the author (Humphreys, 1960) published the intercorrelations of semester grade averages, each semester's average being independent of the rest, and compared their intercorrelations to those of successive trials in learning a motor skill, of successive tests of intelligence, and to the Guttman simplex model. Sharply reduced predictions of senior grades could be inferred from this discussion. More recently Juola (1966) has published correlations between common predictors and independent semester averages that confirm this prediction. The author now has data that make possible the computation of predictor correlations and criterion intercorrelations on large numbers of cases. The results are impressive and the implications for selection of students clear. The data also suggest possible changes in academic dismissal and probation regulations.

[1] The author wishes to acknowledge the very substantial help given him in obtaining and analyzing these data by the University of Illinois Office of Admissions and Records, John Holland and the American College Testing program, and his research assistant, Diane McGrath. The latter was supported by the University Research Board, the University of Illinois.

METHOD

Semester grade-point averages were computed independently for each freshman who entered the University of Illinois in 1962 and 1963 for each semester he was in residence. Eight independent averages were available for the 1962 class and six for the 1963 class. Data on high school rank in class and on the separate tests and composite of the American College Testing (ACT) program were added. Intercorrelations were computed separately for the two sexes and for the college in which the student was enrolled. Finally an aggregate correlational matrix was obtained in which each separate correlation was weighted by the N on which it was based. Thus any between-groups correlations arising from sex or college differences were controlled.

An important consequence of the above procedure was that the Ns vary markedly from Semester 1 to Semester 8. The very large drop from Semester 6 to 7 was the result of starting the study before the end of the senior year for the second of the two classes. A good deal of the change in Ns overall, however, is due to academic selection. Academic dropouts decrease the range of talent and attenuate correlation coefficients. Correction for this attenuation posed a problem. Standard deviations could be computed both for the predictors and for the criteria. The correction formula using ratios of predictor standard deviations assumes that the restriction is due only to a cut on the criterion scores and that there is homoscedasticity in the arrays, defined by the criterion, of the predictor scores. The results from the use of this correction formula are highly sensitive to variations from both assumptions. The formula which uses the ratios of the criterion standard deviations assumes that the cut is on the criterion only, but it is not as sensitive to small variations in the size of the standard deviations. It does assume, however, that the units of measurement are the same for each of the criterion measures. Superficially, different grade-point averages seem based upon the same units of measurment. Closer analysis leads to doubt which cannot be resolved by any empirical test. Do University of Illinois faculty members use the same scale of measurement in assigning grades to

freshmen and to seniors? Does the decrease in the standard deviation of these averages reflect the quality of the academic performance, stereotypes concerning the performance of freshmen and seniors, or some combination of the two?

Because of the doubts concerning corrections for restriction of range of talent it was decided to form a new aggregate correlational matrix in which N was for all practical purposes constant throughout the table. It is possible to do this by selecting graduating seniors rather than entering freshmen. With restriction of range of talent made impossible experimentally, differences in correlations from Semester 1 to Semester 8 cannot be explained by the dropping out of low ability students.

There is, finally, the problem of possible differential reliability of the grade averages from Semester 1 to Semester 8. Split-half estimates would have been possible though laborious to compute. It was decided that the adjacent semester correlations were satisfactory lower bound estimates.

RESULTS

The aggregate table of intercorrelations of predictors and criteria based upon maximum Ns available appears above the diagonal in Table 1. The N for each correlation appears below the diagonal. Standard deviations appear in the diagonal of the table. Correlations between predictors and freshmen criteria have the expected magnitude. Any person who has engaged in college admissions research might well have estimated these values within a couple of hundredths before the study was started. He would probably have done less well, however, in estimating the correlations with senior grades. The latter correlations are certainly nonzero, but the utility of the freshmen predictors for increasing the academic performance of seniors is open to serious question.

The near constancy of the adjacent semester correlations rules out substantial differences in reliability of freshman and senior grades as an explanation of markedly reduced predictor correlations, but the problem of restriction of range of talent remains. This question is partially answered by the data in Table 2 in which N is approximately constant. Standard deviations are again added to the correlational table in the diagonal.

Adjacent semester correlations, the lower bound reliability estimates, are more nearly uniform than before, and the pattern of correlations is very similar to the pattern in Table 1. The intercorrelations of the predictors and the correlations of predictors with early criteria are attenuated as would be expected from the reduced range of talent in the graduating seniors as compared to entering freshmen.

The correlations in Table 2 are helpful in making interpretations of the data, but it is still desirable to have an estimate of what the correlations between predictors and senior grades would have been if there had been no restriction in range of talent. Correlations corrected for restric-

TABLE 1

INTERCORRELATIONS OF HIGH SCHOOL RANK IN CLASS, AMERICAN COLLEGE TESTING PROGRAM VARIABLES, AND GRADE-POINT AVERAGES IN EIGHT SEMESTERS

Item		1	2	3	4	5	6	7	8	9	10	11	12	13	14
High school rank	1	19.56	.383	.427	.307	.311	.446	.513	.424	.324	.302	.270	.264	.242	.216
English	2	8336	3.64	.502	.548	.536	.775	.401	.328	.233	.208	.187	.179	.194	.151
Mathematics	3	8336	8333	5.14	.422	.503	.776	.396	.286	.215	.192	.168	.163	.142	.112
Social science	4	8335	8335	8335	4.72	.604	.803	.365	.308	.239	.215	.204	.215	.155	.129
Natural science	5	8336	8336	8336	8335	4.59	.825	.364	.282	.210	.202	.189	.184	.158	.121
Composite	6	8336	8336	8336	8335	8336	3.63	.475	.376	.284	.261	.240	.236	.203	.160
Semester 1	7	8336	8336	8336	8335	8336	8336	.78	.664	.527	.479	.440	.419	.393	.353
Semester 2	8	7255	7255	7255	7255	7255	7255	7255	.73	.559	.503	.459	.432	.397	.346
Semester 3	9	5352	5352	5352	5352	5352	5352	5352	5262	.68	.588	.516	.482	.452	.363
Semester 4	10	4821	4821	4821	4821	4821	4821	4821	4740	4653	.65	.533	.501	.463	.419
Semester 5	11	4216	4216	4216	4216	4216	4216	4216	4150	4033	4051	.67	.601	.525	.474
Semester 6	12	4077	4077	4077	4077	4077	4077	4077	4014	3900	3873	3920	.63	.566	.505
Semester 7	13	1800	1800	1800	1800	1800	1800	1800	1776	1711	1705	1692	1726	.62	.566
Semester 8	14	1709	1709	1709	1709	1709	1709	1709	1685	1629	1620	1610	1639	1664	.60

Note.—N is maximum for each correlation.

TABLE 2

Intercorrelations of High School Rank in Class, American College Testing Program Variables and Grade-Point Averages in Eight Semesters

Item		1	2	3	4	5	6	7	8	9	10	11	12	13	14
High school rank	1	15.45	.356	.368	.244	.291	.393	.387	.341	.278	.270	.240	.256	.240	.222
English	2		3.32	.485	.545	.542	.773	.345	.278	.226	.236	.236	.222	.216	.160
Mathematics	3			4.95	.395	.471	.764	.279	.189	.171	.171	.145	.162	.156	.121
Social science	4				4.33	.637	.802	.279	.244	.188	.198	.210	.225	.174	.149
Natural science	5					4.21	.829	.306	.255	.177	.200	.184	.202	.159	.126
Composite	6						3.35	.375	.298	.237	.255	.238	.252	.219	.173
Semester 1	7							.57	.556	.456	.439	.399	.415	.387	.342
Semester 2	8								.57	.490	.445	.418	.383	.364	.339
Semester 3	9									.57	.562	.496	.456	.445	.354
Semester 4	10										.54	.512	.469	.442	.416
Semester 5	11											.59	.551	.500	.453
Semester 6	12												.58	.544	.482
Semester 7	13													.59	.541
Semester 8	14														.59

Note.—N is approximately 1,600 for each correlation.

tion of range of talent utilizing the ratios of standard deviations of the criteria in each of the eight semesters are presented in Table 3. (Only two decimal places are retained for corrected values.) It is perhaps a little reassuring with respect to the corrected values to note the constant size of the standard deviations for the criterion variables in Table 2. Since there cannot have been any change in range of talent, the standard deviations should be identical if the scales of measurement used are identical.

An internal check on the accuracy of the corrections for restriction of range is also available. The obtained values in Table 2 can be used to estimate the relationships in Table 1. The comparison between estimated and obtained correlations tests the adequacy of the formula for these data. A good fit leads to greater confidence in the Table 3 results. Such a fit is indicated by the mean algebraic discrepancy between obtained values in Table 1 and values estimated from Table 2 of −.003. The mean absolute discrepancy is only .023.

The reduction in validity in Table 3 as a function of semester in college is dramatic. Common variance between high school rank in class and college grades changes from 26% in Semester 1 to about 8% in Semester 8. The composite score from the academic ability test and each of its components show similar patterns although at a somewhat lower level of common variance. The data suggest that people are changing and that "aptitude for college work" is far from stable.

It would be desirable to have retest data on the ACT at the end of the senior year in order to pin down the explanation phrased in terms of change in people since there may be other logically permissible explanations of the present data. Nevertheless, the phenomenon is so general (Humphreys, 1960, 1967) and in the present data so consistent among colleges and sexes, for example, from males in the homogeneous College of Engineering to either males or females in the heterogeneous College of Liberal Arts and Sciences,

TABLE 3

Correlations Corrected for Restriction of Range of Talent between Predictors and Criteria

Item	Semester							
	1	2	3	4	5	6	7	8
High school rank	.513	.45	.37	.36	.31	.32	.30	.28
English	.401	.35	.27	.25	.22	.22	.24	.20
Mathematics	.396	.30	.25	.23	.20	.20	.18	.15
Social science	.365	.33	.27	.26	.24	.27	.19	.17
Natural science	.364	.30	.24	.23	.22	.23	.20	.16
Composite	.474	.40	.32	.31	.28	.29	.25	.21

that logical alternatives appear to be low probability alternatives.

DISCUSSION

Is the change in intellectual performance during the college years the result of continuing maturation which proceeds at a differential rate from one person to another or is the change due to the stimulation and viscissitudes of the college environment? An overall increase in intellectual ability, which would suggest growth or maturation, would be relevant, but there is no direct evidence concerning this. It might be argued that the kinds of performance measured by an intelligence test such as the Stanford-Binet do not show much increase after 16. On the other hand, there is every reason to believe that students are learning something intellectual while in college and that certain kinds of intellectual performance are increasing in level. Since it is difficult to support any fundamental difference between aptitude and achievement (Humphreys, 1962) it can be assumed that there have been changes in the intellectual level of the group as a whole in addition to the changes in the rank ordering of individuals that were observed. There may well be both biological growth factors and environmental stimulation and deprivation factors involved, but from certain points of view an explanation is immaterial. The empirical fact remains that there is a good deal of instability in intellectual performance during the 4-year undergraduate period and as a result the correlations of predictors with criteria show a great deal of shrinkage over this period of time. Senior performance is not predicted well enough from freshman information for one to be at all content with present college admission practices.

The author has no better selection instruments to recommend than the present ones. These data do suggest, however, some altered approaches to admissions research. Motivational and interest variables have not shown much promise as predictors of freshmen grades. It is probable, however, that so much of the reliable variance in freshmen grades is associated with high school senior academic performance, which is measured both by rank in class and by the ACT, that there is little left to be associated with nonintellectual variables. A difference score or a residual score computed from freshmen and senior grades would have sufficient reliability to be predictable and would be an appropriate criterion against which to validate nonintellectual variables.

A very different approach to selection research may be sounder theoretically and practically than the use of grades or a criterion derived from grades. Perhaps admission tests should be validated primarily against staying in college versus dropping out. While this criterion would currently in most colleges be heavily contaminated with freshman academic performance, this is not necessary. It is argued below that the data presented here suggest changes in failing and probation regulations. Such changes, if made, would have a substantial effect on the nature of the dropout criterion and would allow colleges to retain a substantial number of potentially baccalaureate-level students.

There is indeed ample basis for discontent with most procedures concerned with placing students on probation and dropping students from college for academic deficiencies. It is obvious that a good many students who are dropped at the end of the first semester would do acceptable work later in college. Students are also placed on unrealistic probation requirements (for example, the lower the academic performance, the higher the requirement the next term) and are subsequently dropped, who would be able to do acceptable work in later years. There seem to be two implicit assumptions used in establishing failing and probation regulations. One concerns ability, the other motivation. If a student has the necessary ability, probation should put him on his mettle, and he will come through if he works. If he works and does not come through, he did not have sufficient innate ability. If he seems, by some measure, to have the necessary innate ability but does not come through, he did not work.

In place of these all or nothing assumptions about ability and motivation, a

radically different interpretation should be attempted. The first supposition will be that intelligence is not fixed, that there is no measurable or inferrable (from measurements) innate capacity, and that gain in intellectual functioning continues indefinitely with adequate stimulation in a healthy organism. By the same token loss occurs without adequate stimulation. The second supposition will be that intellectual functioning, either on psychological tests or in the classroom, depends on a very broad, cumulative, well-learned repertoire of skills, knowledge, modes of performance, etc. Furthermore, this repertoire increases with age and experience. Also for a given absolute amount of change, the relative change on which test-retest correlations depend is smaller when the base is large than when it is small; that is, the amount of change which is of interest here is a function of age and experience. It would take considerable time for a student to gain enough to change appreciably his rank order in his peer group; the amount of time required to make a change of a given magnitude would be a function of age. For very young groups time would be measured in months, for older groups in years. A dull student can change into a bright student, but this always happens gradually and the rate of change is slower among college students than among grade-school students.[2]

It would also be expected that some specialized forms of intellectual functioning are less dependent on the total accumulation of intellectual skills and knowledge than others. The rank order of older, more mature students may change more rapidly in certain kinds of learning situations than in others. Although quite far afield, it might be noted that change in rank order of performance in a discrimination-reaction time task takes place very rapidly (Fleishman, 1955). Holding age constant, changes in rank order on the Stanford-Binet which taps very general, old, well established learning would be expected to proceed more slowly than changes in rank order in learning a foreign language. College specialization would generally represent a learning situation that did not depend on the total accumulation as much as does performance on the Binet. Thus test-retest correlations on the Binet should show more stability than similar correlations for the ACT, but this would reflect the age and generality of the learning represented, not a difference between aptitude and achievement.

On the basis of the present data and of the preceding theorizing it would be desirable to establish probation regulations such that the student would be kept in school as long as he was making minimally adequate progress toward an acceptable graduation average. The goal, in a nutshell, is to give him time to change his level of performance. One simply cannot predict well enough from freshman academic deficiency to senior performance. The typical counterargument is that the colleges should get rid of marginal students and make space available for those of better quality. Because of the instability of performance, however, definition of better quality is just as suspect as marginal quality.

There is also an economic argument to suggest changes in probation and failing regulations. Once the student has invested a semester in college study, and once the college has invested in the student by admitting him, housing him, and teaching him, if he can attain graduation standards on or near schedule, the social gain is maximized, and the expense minimized, by retaining him. This approach, incidentally, should not be termed "coddling," and it need not deteriorate into coddling if graduation standards are maintained.

The problem of admissions research and the choice of a criterion may now be discussed. Regulations which would allow a student time to overcome initial academic

[2] There are also biological factors involved in intellectual learning that are completely neglected by this discussion. Thus the time required to change the rank order of a student by some given amount is also a function of the biological organism involved. Since there are no measurement operations that allow the assessment of these biological factors independently in the human, and since these factors are not within social control, they can be disregarded in this discussion.

deficiencies would indeed change the nature of the dropout criterion. It would become a better measure of the "sticking to the task" trait and less highly related to initial academic performance. It would allow for more covariance between nonacademic predictors and the criterion. It presents, however, a research task of major proportions: the finding of good nonacademic predictors.

REFERENCES

FLEISHMAN, E. A., & HEMPEL, W. E. The relation in a visual discrimination reaction task between abilities and improvement with practice. *Journal of Experimental Psychology*, 1955, **59**, 301-310.

HUMPHREYS, L. G. Investigations of the simplex. *Psychometrika*, 1960, **25**, 313-323.

HUMPHREYS, L. G. The nature and organization of human abilities. *19th Yearbook of the National Council on Measurement in Education*, 1962, 39-45.

HUMPHREYS, L. G. Problems in personnel research. In A. L. Fortura (Ed.), *Personnel research and systems advancement*. Lackland Air Force Base, Texas: Personnel Research Laboratory, 1967.

JOULA, A. E. Prediction of successive terms performance in college from tests and grades. *American Educational Research Journal*, 1966, **3**, 191-197.

(Received November 3, 1967)

Educational Psychology

SECOND EDITION

LEE J. CRONBACH, Stanford University

This outstanding textbook strongly emphasizes intellectual learning and development and provides a full presentation of the psychological theory and research that underlie educational practice. Two separate Test Item Files are available. 706 pages. $9.25

STUDENT GUIDE WITH PROGRAMED UNITS

prepared by GABRIEL M. DELLA-PIANA.

Paperbound. 151 pages. $2.45

READINGS FOR EDUCATIONAL PSYCHOLOGY

Prepared by ELLIS BATTEN PAGE
In consultation with LEE J. CRONBACH

Paperbound. 404 pages. $4.95

"I have long regarded Cronbach's first edition as the finest text available in the area of educational psychology and the second is a definite improvement. Page's selection of readings is well designed to highlight issues in the field. Both volumes are a contribution of inestimable value."

—PHILIP EDDY, University of Massachusetts

Learning

J. CHARLES JONES, Bucknell University

A concise interpretation of recent psychological research findings on learning which provides the prospective and practicing teacher with a theoretical basis for examining procedures and problems of education.

Paperbound. 179 pages. $2.45

"Quite comprehensive in its coverage and quite readable, Learning *should prove to be a very worthwhile resource for both teachers and counselors."*

—LYLE L. MILLER, University of Wyoming

A Social Psychological View of Education

CARL W. BACKMAN and PAUL F. SECORD, University of Nevada

Presents a clear and well-documented review of the concepts and data from social psychology that are directly pertinent to the role of the teacher in the classroom, in the school system, and in society.

Paperbound. 152 pages. $2.45

"The presentation of the information in this volume is succinct and up to date, giving the students an opportunity to consider additional facets in the evolution of their educational philosophy."

—EDWIN I. MARIENHOFF
Hunter College of The City University of New York

HARCOURT, BRACE & WORLD, INC.

New York / Chicago / San Francisco / Atlanta

ALLYN AND BACON, INC.

1968-HUMAN DYNAMICS IN PSYCHOLOGY AND EDUCATION

Don E. Hamachek, Michigan State University. This selection of readings deals not only with learning, motivation, intelligence, etc., but also with such topics as innovations in instruction, the disadvantaged, personal adjustment, and understanding one's self. Frequently, opposing views are presented. 1968. Paperbound. 695 pp.

HUMAN DEVELOPMENT IN WESTERN CULTURE, Second Edition

Harold W. Bernard, Oregon State System of Higher Education. Organized around the continuity of developmental tasks and the institutions which influence development through childhood and adolescence. Concludes with a chapter on individual self-realization. 1966. 552 pp.

READINGS IN HUMAN DEVELOPMENT

Harold W. Bernard and **Wesley C. Huckins,** both of the Oregon State System of Higher Education. Represents the new emphases in human development, such as the implications of social class, the meaning and impact of the war on poverty, and the need for preventing school drop-outs. 1967. Paperbound. 470 pp.

EVALUATING PUPIL GROWTH, Third Edition

J. Stanley Ahmann, Colorado State University; and **Marvin D. Glock,** Cornell University. Principles of Tests and Measurement. Presents a comprehensive discussion of the problems and techniques of evaluation. The discussion of standardized instruments has been modernized to include, for example, the Differential Aptitude Test. 1967. 605 pp.

1968-EDUCATIONAL PSYCHOLOGY, Third Edition

James M. Sawrey and **Charles W. Telford,** both of San Jose State College. The strong emphasis on learning and motivation, characteristic of the previous editions, is carried on in this one. A new chapter on the educational significance of individual differences includes discussion of the educational provisions and practices in relation to the intellectually, culturally, socially or physically deviant individual. Coverage of the school as a psychosociological unit is greater, as is consideration of creativity. 1968. 630 pp.

1968-STUDENT GUIDE

to accompany Educational Psychology, Third Edition. **Earl L. Jandron,** San Jose State College. 1968. Paperbound. 194 pp.

READINGS IN EDUCATIONAL PSYCHOLOGY

Ronald K. Parker, The Florida State University. Designed to accompany Educational Psychology, Third Edition or any other basic text. A current balanced selection of articles, providing the undergraduate with a frame of reference for appreciating and understanding research in the field. 1968. Paperbound. 608 pp.

470 Atlantic Avenue, Boston, Ma 02210

Education 1969

forthcoming books from HRW

School Learning:

An Introduction to Educational Psychology

By David P. Ausubel, The City University of New York, and

Floyd G. Robinson, The Ontario Institute for Studies in Education

In the undergraduate-level, classroom-tested version of David P. Ausubel's *Educational Psychology: A Cognitive View*, the authors explore only the actual kinds of learning that take place in the school: meaningful verbal learning and subject matter learning. They begin with a self-consistent theory of classroom learning which is specifically appropriate and relevant for such learning, rather than a theory which is eclectic or directed toward all kinds of learning. The relevance of this theory for school learning is demonstrated by hundreds of examples of classroom teaching.

April 1969 576 pp. $8.50 (tent.)

Instructor's Manual
Student Workbook

The Research Process in Education

By David J. Fox, The City College of the City University of New York

Through a comprehensive explanation of each stage in the research process, Professor Fox gives the student a complete working knowledge of the methods, techniques, and statistical concepts necessary to design and carry out a research project. To introduce students to the research process, the author begins with a case study of a complete project and then develops a model of the stages of research with exercises that can be used by the student to design his own research.

April 1969 672 pp. $8.95 (tent.)

Standardized Tests in Education

By William A. Mehrens and Irvin J. Lehmann, both of Michigan State University

Classroom teachers, counselors, and school administrators will find all the information needed to select, administer, and interpret standardized tests correctly in this volume. The authors analyze, classify, evaluate, compare, and contrast available tests and demonstrate how to interpret the information presented in test manuals.

January 1969 320 pp. $4.95 paper (tent.)

Holt, Rinehart and Winston, Inc.

383 Madison Avenue, New York, New York 10017

Educational and Psychological Measurement

David A. Payne and Robert F. McMorris

For courses focusing on tests and measurements, this text contains fifty-four theoretical and research papers providing a broad perspective on test development with an emphasis on the assessment and prediction of learning outcomes. A manual of examination questions is available. *1967. 419 pages. Paper. $5.75*

The Specification and Measurement of Learning Outcomes

David A. Payne

Intended for undergraduate and graduate courses, the text has a single purpose: to provide the classroom teacher with an efficient and practical set of techniques to aid in evaluating student achievement. *1968. 209 pages. Paper. $2.75*

Studies in Educational Psychology

Raymond G. Kuhlen

For courses focusing on the psychology of school learning, this text contains fifty papers dealing with the learning process with special attention to the cognitive, personality, and motivational factors involved. *1968. 482 pages. Paper. $5.75*

Readings in the Psychology of Childhood and Adolescence

William J. Meyer

For courses focusing on human development, this book contains sixty-five theoretical papers and research reports dealing with the main features of child and adolescent development. *1967. 436 pages. Paper. $5.25*

The Psychology of Vocational Choice

John L. Holland

This text offers a concise synthesis of current knowledge of vocational adjustment, and further presents original theory and suggestions for new research. Written primarily for students and professional audiences, the book effectively interprets data for practical application. *1966. 132 pages. Paper. $1.95*

Methods of Studying the Individual Child: The Psychological Case Study

John Rothney

Intended for teachers, counselors and school psychologists in training and in service, this text advocates the case study as a primary aid in understanding and furthering the development of the individual child. *1968. 144 pages. Paper. $1.95*

BLAISDELL PUBLISHING COMPANY

A Division of Ginn and Company

275 Wyman Street
Waltham, Massachusetts 02154

Educational Psychology

An Application of Social and Behavioral Theory

LOUIS M. SMITH *and* BRYCE B. HUDGINS, *both of Washington University at St. Louis*
Knopf; 576 pages; $8.95; illustrated

An approach to teaching that is both practical and theoretical. Drawing from the sciences and the humanities, Professors Smith and Hudgins offer detailed analytical and critical discussions of such questions as concept learning, creativity and influence technique while also examining traditional subject matter. "An excellent text! I am particularly appreciative of the incorporation of material from social and behavioral theory."—Sidney Yudin, *New York University*

Individualizing Instruction in the Elementary School

GEORGE I. THOMAS, *State Department of Education, New York and* JOSEPH CRESCIMBENI, *Jacksonville University*
Random House, 1967; 448 pages; $7.50

A compendium of techniques which emphasizes the individualizing approach to teaching. This historical-theoretical overview of elementary education is followed by a detailed subject-by-subject treatment with illustrations of concrete situations so that teachers can adapt to their own particular needs a vast array of techniques.

Guiding the Gifted Child

GEORGE I. THOMAS *and* JOSEPH CRESCIMBENI
Random House; 224 pages; $2.25 paperbound

A study of gifted children—their classroom needs, expectations and problems. The text confronts the problem of recognition and presents a plan for meeting the needs of gifted pupils through acceleration and enrichment.

Mental Health in the Schools

THOMAS A. RINGNESS, *University of Wisconsin*
Random House, 1967; 500 pages; $6.95

A scientifically grounded and practically oriented book on "mental health." Professor Ringness has not confined himself to a single rigid theory. Instead, he has found an eclectic approach based primarily on the roles of reinforcement and learning in personality development.

Now in paperback . . .

Six Psychological Studies

JEAN PIAGET
A Vintage Book, September 1968; 192 pages; $1.65
A useful general introduction to the work of the eminent Swiss psychologist.

BORZOI College Textbooks

RANDOM HOUSE, Inc. ALFRED A. KNOPF, Inc.
The College Department
501 Madison Avenue New York 10022

Consider this important new text...

CASEBOOK ON ETHICAL STANDARDS FOR PSYCHOLOGISTS

The application of the 19 principles comprising the Code of Ethics for all psychologists is presented. Each principle is illustrated by actual, but disguised, cases involving alleged violations of the principle and action taken by the Board of Professional Affairs.

Four Appendices state and amplify the additional principles applicable to specific areas of psychology:

—*Ethical Practices in Industrial Psychology*

—*Automated Test Scoring and Interpretation Practices*

—*Standards for APA Directory Listings of Private Practice*

—*Guiding Principles for the Humane Care and Use of Animals*

The arrangement of the Casebook into a general section of cases and four additional specific sections increases its versatility as a reference book for individuals and as a text for instructional purposes.

Price: $1.00

Order From:

AMERICAN PSYCHOLOGICAL ASSOCIATION

1200 Seventeenth Street, N. W.
Washington, D. C. 20036

McGraw-Hill Books in Educational Psychology

Coming Soon From McGraw-Hill

PSYCHOLOGICAL FOUNDATIONS OF LEARNING AND TEACHING

John A. R. Wilson, Professor of Education, University of California; **Mildred C. Robeck,** Professor of Education, University of Oregon; and **William B. Michael,** Professor of Education and Psychology, University of California at Santa Barbara. *Available Winter, 1969.*

The purpose of this book is to present a theoretical model by which a teacher may structure his course to encourage students to higher levels of motivation and performance, stressing the uniqueness of the individual and his problems.

PSYCHO-SOCIAL PRINCIPLES APPLIED TO CLASSROOM TEACHING

William F. White, Associate Professor, The University of Georgia. *Available Spring, 1969.*

This text provides a social psychological approach to classroom teaching to bring depth to the teacher-pupil relationship, emphasizing research substantiating these principles.

For the Fall Term ....

READINGS IN PSYCHOLOGICAL FOUNDATIONS OF EDUCATION

Walter H. MacGinitie; and Sam Bell, Editors, both of Teachers College, Columbia University. *Off Press.*

Intended for courses in Educational Psychology, the articles selected were chosen for their substance and effect on education or their usefulness in showing methodology and techniques in educational and psychological research.

HUMAN MEMORY

Jack A. Adams, University of Illinois. *Off Press.*

First book in the field of memory research to relate the theories of memory to scientific data gathered from the laboratory work of experimental scientists.

McGRAW-HILL Book Company

330 West 42nd Street,
New York, New York 10036

JOURNAL OF
EDUCATIONAL PSYCHOLOGY

Volume 59 | December 1968 | Number 6, Part 1

CONTENTS

This is the last issue of Volume 59.
Volume Title Page and Contents of the Volume appear herein.

© 1968 by the American Psychological Association, Inc.

Journal of Educational Psychology
1968, Vol. 59, No. 6, 381–387

PROGRAMMED INTRODUCTION TO PSYCHOLOGY VERSUS TEXT-BOOK STYLE SUMMARY OF THE SAME LESSON

MARIANNE RODERICK AND RICHARD C. ANDERSON[1]
University of Illinois

Either the 1st 4 sets of the Holland-Skinner (1961) program or a summary of the material contained in the program was given to 85 college undergraduates and 116 high school seniors. Overall, those who completed the program scored higher on the achievement test than those who studied the summary. The advantage of the program was greatest (*a*) with high school rather than college students; (*b*) on the delayed rather than the immediate achievement test; (*c*) on short-answer rather than multiple-choice test items. Among college undergraduates, for whom the program was designed, the results failed to show better achievement for the program than for the summary, but the program took 4 times as long to complete.

Critics of programmed instruction often voice the complaint that programs present material in steps that are unnecessarily small, that they involve too much repetition, and that such features are not required to produce learning with sophisticated students. Pressey and Kinzer (1964) have completed a study that gives empirical support to doubts about the efficiency of small-step programs. They prepared a succinct, textbook-style summary of the first two sets of *The Analysis of Behavior* (Holland & Skinner, 1961). The summary consisted of 643 words, while the section of program upon which it was based contained 1,710 words and also entailed 84 written responses. Students took eight times as long to complete the program as they did to read the summary, yet those who received the summary scored higher on the posttest. Students who completed nine "auto-elucidative" questions in addition to reading the summary obtained the highest posttest scores of all. The Pressey and Kinzer experiment suffered from methodological shortcomings. Instead of random assignment of *S*s to treatments, whole classes received one treatment or another. The posttest consisted of an essay examination which was described as "carefully graded" but about which no further information was provided.

The results of the Pressey and Kinzer study would appear to be inconsistent with the findings of other research that has employed *The Analysis of Behavior.* One time-consuming feature of this program is the frequent requirement to make written responses. Yet Williams (1963) has found that overt responding produces better achievement than reading the program with filled blanks. Casual inspection of *The Analysis of Behavior* suggests that it is a redundant program. It contains many groups of frames in which equivalent responses are required in the presence of identical or nearly identical stimuli. Herein lies another possible contributor to the inefficiency which Pressey and Kinzer seem to have found. However, Coulson and Silberman (1960) reduced a 104-frame section of *The Analysis of Behavior* to 56 frames by removing frames judged to be redundant. The *S*s who received the standard program scored higher on the posttest than those who received the shortened version.

It would be easy to discount the Pressey and Kinzer study because of its shortcomings and the contradictory findings from other experiments. Still, it does seem possible that redundancy and overt responding are necessary to produce satisfactory achievement given, and only given, the constraints of a small-step pro-

[1] The authors are indebted to Thomas Anderson, Gerald Faust, and Philip Zediker for assistance in running *S*s and to Elwood Leslie for help in machine-scoring tests and completing item analyses. They are grateful to the McGraw-Hill Book Co. for permission to reproduce sections of *The Analysis of Behavior*.

Copyright © by the American Psychological Association, Inc.

gram and that these time consuming features are not necessary to attain satisfactory achievement from a text. In other words, the value of such features may depend upon the form of the material in which they are included.

There is a plausible argument for considerable redundancy. The presumption is that many apparently similar encounters with the material are necessary in order to arrange discriminations among the terms and concepts being taught. A single statement of a principle may be sufficient if one's goal is merely to have the student name the principle when it appears in the verbatim form employed during instruction. If, on the other hand, one wants the student to be able to recognize various expressions of the principle, to discuss the principle fluently in his own words, to identify new instances of the principle, and to apply the principle to novel cases not treated during the course of instruction, then it may well be necessary to require the student to deal with a variety of forms of the principle and a variety of examples.

There is also a plausible argument for requiring the student to make overt, constructed responses. People learn what they are led to do. A person may spontaneously make appropriate covert responses when reading a text. But then again he may skim, skip difficult sections, or render the material in a way different from that intended by the author. The argument is that the requirement to make overt responses helps to ensure that the student will actually make the responses necessary for learning.

It remains to be seen whether the theoretical advantages of redundancy and overt responding are obtained in practice. To a greater or lesser degree, depending upon the task, the student will already be capable of the responses and discriminations entailed in a lesson. Because of what he has previously learned, the student sometimes may almost spontaneously generalize to appropriate stimulus and response classes. These entering behaviors may be systematically undervalued by programmers who have been exhorted to use small steps, keep error rate low, and leave nothing to chance. Students may often be compelled to endure a lengthy program when several pages of clear English could evoke the desired performance. Commonly employed techniques for the development and validation of programs do not protect against inefficiency. Presumably during the course of tryout and revision the programmer will discover the instances in which he has underestimated the difficulty of teaching a concept. But what of the instances in which difficulty has been overestimated? As Markle (1967) has indicated in her excellent analysis of the problem, frames upon which students make few errors are very unlikely to be eliminated from a program. Nor will sections of a program be compressed when students do very well on criterion test items measuring what these sections teach. Programs inevitably grow longer rather than shorter when revised. There is no empirical technique in use to detect superfluous redundancy.

If programmers tend to underestimate entering behavior in the intended target population, an empirical demonstration of the value of considerable redundancy and overt responding would be difficult. However, a lesson characterized by a gradual progression of small steps, repetition and review, and the requirement to make overt, constructed responses might show to advantage in a population less skilled than the intended target population. One purpose of the present experiment was to compare a small-step program and a textbook-style summary of the material taught by the program with students who were grossly deficient in the entering skills manifested within the population for which the program was designed. These students were assumed to have available few of the responses to be acquired and were assumed unlikely to discriminate and generalize spontaneously in an appropriate manner among stimuli and responses. Consequently, the program was expected to work much better (though not necessarily well in abosolute terms) than the summary for the students relatively deficient in entering behavior but perhaps only slightly better than the summary for students

from the target population. The program used in this study, *The Analysis of Behavior,* was designed for use with college sutdents. The program and summary were compared with both college and high school students.

The redundancy in many small-step programs may not be necessary to produce adequate performance on an immediate test. However, it is well established that repetition and spaced review facilitate retention. Another purpose of the study reported herein was to compare a program and a succinct summary on both an immediate and a delayed achievement test. The program was expected to show to greater advantage on the delayed test than on the immediate test.

The final purpose of the present experiment was to compare a program and a summary on both short-answer test items and equivalent multiple-choice test items. The student may be adequately prepared to recognize a new technical term if he has simply read a passage within which the term was defined and illustrated. However if he is expected to be able to produce the new term, there is good reason to believe that the requirement to produce the term during the course of instruction will be helpful.

METHOD

Subjects and Experimental Design

Eighty-five college sophomores, juniors, and seniors enrolled in an introductory course in educational psychology and a heterogeneous group of 116 high school seniors served as *S*s. One college *S* was dropped due to failure to complete the program, while five high school and seven college *S*s were lost because they were absent for the delayed achievement test.

A 2 × 2 × 2 × 2 design was employed with a repeated measure defining one of the factors. The first factor was training method. The *S*s completed either the program or the summary. The second factor was *S* status. The *S*s were either high school seniors or college undergraduates. Retention interval was the third factor. The final factor was test mode. Both a short-answer and a multiple-choice achievement test were given to all *S*s.

Instructional Materials

About half of the *S*s received the standard version of the first four sets of *The Analysis of Behavior.* Each frame occupied a 3⅝ × 8½ inch page; the answer to that frame appeared at the top of the following page. The program was mimeographed on blue paper through which the following page could not be read. Each set was stapled along the left margin to form a separate booklet.

The remaining half of the *S*s studied a textbook-style summary of the first four sets of *The Analysis of Behavior.* The summary was initially written by "lifting" material from the program in the order in which it appeared there. The material was later condensed and arranged in paragraph form to make it readable. Technical terms were underlined upon their introduction into the text, but not again. One example was used to illustrate each principle; no other redundancy was included. Several psychologists read the summary and made suggestions for its improvement. All agreed that material coverage was adequate. Prior to the experiment about 50 high school seniors, attending a different school from the one in which the experiment was conducted, and about 25 college undergraduates, enrolled in an introductory educational psychology course, completed the achievement test on an "open summary" basis. They were asked to read the summary and then take the test, searching through the summary to find or verify answers. From 31% to 100% of the students answered each question correctly. The lowest percentages were obtained from the high school seniors on several short-answer items. However, none of the latter items was answered correctly by any of the high school seniors in another group of 50 which was not exposed to the summary. These data indicate that all of the test items could be answered on the basis of material contained in the summary.

The final version of the summary contained 1,799 words.[2] The program contained 3,398 words and involved 142 written responses.

Procedure

The *S*s were assigned to treatments by issuing them "tickets" from a deck stacked in a predetermined random order. The ticket directed *S* either to a room in which the summary was employed or to a different room in which the program was used. The program and summary were not employed within a single room because the program takes more time. There could have been a reactive effect had those completing the program seen many others finishing early.

The *S*s in the program group received directions similar to the standard program directions published in *The Analysis of Behavior.* Those who received the summary were told to read at their

[2] For copies of the summary and the achievement test, order NAPS Document No. 00066 from ASIS National Auxiliary Publications Service, % CCM Information Sciences, Inc., 22 West 34th Street, New York, New York 10001; remitting $1.00 for microfiche or $3.00 for photocopies.

own rate and were expressly permitted to reread all or part of the material if they so desired.

All *S*s were told that they would receive a test when they finished the program or summary. Assistants completed a control sheet upon which were recorded the order in which *S*s completed the treatment and the time required for each to do so. As each pair of *S*s finished the program or summary one of the two was randomly assigned to receive the immediate achievement test while the other received an irrelevant (verbal reasoning) test as a time filler and placebo. This procedure equated the immediate and delayed achievement test groups in terms of training time.

High school *S*s received the delayed achievement test 7 days after the immediate test. The interval between the two tests ranged from 6 to 9 days for the college *S*s. The delayed test was not announced and the teachers and others cooperating in the experiment were asked not to reveal that a test would be given again. The delayed test was given at a regularly scheduled meeting of an educational psychology course in the case of college *S*s. In the case of the high school *S*s assistants made what was presumably an unexpected visit to the cooperating school to administer the delayed test. There is no evidence that *S*s expected a second test. On the contrary, many *S*s seemed genuinely surprised when the delayed test was administered.

The same measure, consisting of 19 short-answer items and 19 equivalent multiple-choice items,[3] was used as both an immediate and delayed test. For pairs of short-answer and multiple-choice items, the item stems were identical, or nearly so. In 20 of the test items, the wording of the items was essentially the same as the wording of statements included within the lesson materials and/or the items involved the same examples as were used to illustrate concepts or principles within the lesson materials. The remaining 18 items contained wording substantially different from any instructional statement and/or the items entailed new examples not included within the lesson materials. On each occasion the short-answer section of the test was administered first and collected before the multiple-choice section was distributed. The short-answer section was scored on the basis of a criterion list of acceptable answers. The multiple-choice section was machine scored and corrected for guessing.

RESULTS

Table 1 contains the achievement test means for the various experimental conditions. Since there were disproportionate numbers of cases per cell, an unweighted means analysis of variance, summarized in Table 2, was performed. This analysis did not include the delayed achievement test scores of *S*s who had also completed the immediate test, but only the delayed test scores of *S*s who had received an irrelevant test immediately after the treatment and who were therefore taking the achievement test for the first time.

All four main effects were significant. The unweighted mean percent correct on the achievement test was 61.7 for those who completed the program and 54.4 for those who read the summary, 39.1 for high school students and 77.1 for college

TABLE 1

MEAN PERCENT CORRECT ON THE ACHIEVEMENT TEST

Condition	High school		College	
	Immediate	Delayed	Immediate	Delayed
Program				
N	30	28	19	19
SA	48.1%	40.0%	83.9%	73.1%
MC	45.3%	41.5%	83.9%	77.8%
Summary				
N	28	25	22	17
SA	33.3%	29.3%	84.4%	57.6%
MC	38.3%	36.8%	84.0%	71.5%

Note.—Included are delayed test scores of only *S*s who were taking the achievement test for the first time. Abbreviated: SA = short answer, MC = multiple choice.

TABLE 2

ANALYSIS OF ACHIEVEMENT TEST VARIANCE

Source	*df*	*MS*	*F*
Between *S*s			
Training method (M)	1	87.18	5.66*
Subject status (S)	1	2349.82	152.48**
Retention interval (R)	1	137.54	8.92**
M × S	1	6.42	.42
M × R	1	6.67	.43
S × R	1	38.65	2.51
M × S × R	1	20.93	1.36
*S*s within groups	180	15.41	
Within *S*s			
Test mode (T)	1	44.38	8.32**
M × T	1	26.26	4.92*
S × T	1	2.36	.44
R × T	1	34.26	6.43*
M × S × T	1	1.38	.26
M × R × T	1	3.17	.60
S × R × T	1	7.71	1.45
M × S × R × T	1	6.75	1.27
*S*s Within Groups × T	180	5.33	

Note.—The analysis involved the delayed test scores of only *S*s who were taking the achievement test for the first time. The analysis was completed before mean scores were converted to percentages.

* $p < .05$.

** $p < .01$.

[3] There were six additional multiple-choice items for which there were no matching short-anwer items. The results with these six items are not reported herein, though these results did parallel those obtained with the rest of the test.

students, 62.6 when the test was given immediately and 53.5 when it was delayed, and 56.2 on the short-answer items and 59.9 on the multiple-choice items.

There were two significant interactions. The Training Method × Test Mode interaction is graphed in Figure 1. The figure shows that the advantage of the program over the summary was greater on short-answer items than it was on multiple-choice items. Figure 2 pictures the Retention Interval × Test Mode interaction. There was a smaller decrement over the retention interval on multiple-choice items than on short-answer items.

Completed also was a second analysis involving only delayed achievement test scores. The new variable of interest, which turned out to make a significant difference ($F = 8.68$, $df = 1/166$, $p < .01$), was whether *S* had received the immediate achievement test or an irrelevant test in its place. Those who received the immediate achievement test showed an unweighted mean percent correct of 63.5 on the delayed achievement test while the percent correct for those who received the irrelevant immediate test was 53.5. Whether or not *S* took the immediate achievement test interacted significantly with test mode ($F = 9.75$, $df = 1/166$, $p < .01$). Taking the immediate achievement test (see Figure

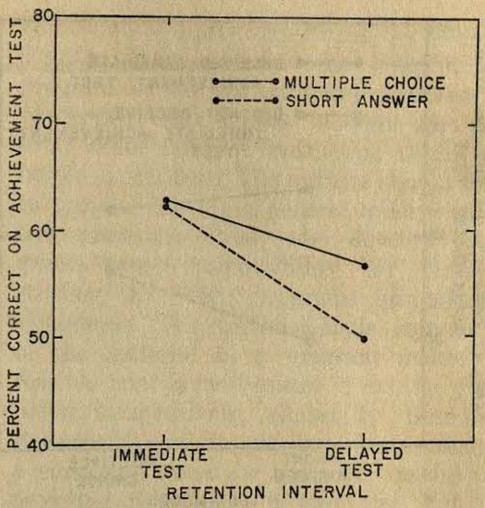


FIG. 2. Percent correct on short-answer and multiple-choice items as a function of retention interval.

3) had a greater effect on performance on short-answer items than on multiple-choice items contained in the delayed test.

Table 3 contains mean training times. Notice that Ss spent about five times as long to complete the program as they did to read the summary.

DISCUSSION

Like Pressey and Kinzer (1964), the present authors found that college undergraduates who complete the initial sections of *The Analysis of Behavior* score no higher on an achievement test given immediately (83.9%) than do undergraduates who study a summary of the material contained in the program (84.2%). Furthermore, in the present experiment undergraduates spent about four times as long working on the program as they did reading the summary, once again approximately replicating Pressey and Kinzer. However, unlike the Pressey and Kinzer study, which was limited to the performance of college undergraduates on an immediate achievement test, the present study showed a significant overall achievement advantage for the program.

It was expected, for reasons outlined earlier, that the program would be most

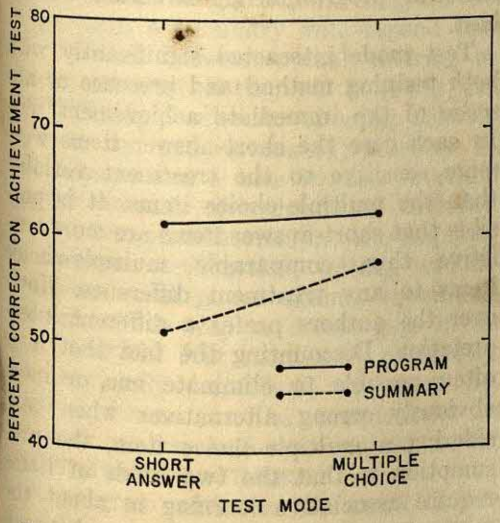


FIG. 1. Percent correct on the achievement test for groups that received the program or the summary as a function of test mode.

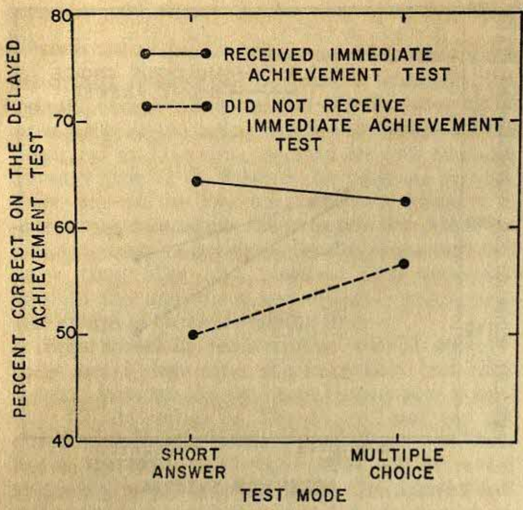


FIG. 3. Percent correct on the delayed achievement test for groups that did or did not receive the immediate achievement test as a function of test mode.

markedly superior to the summary: (*a*) with high school rather than college students; (*b*) on the delayed rather than the immediate achievement test; and (*c*) on short-answer rather than multiple-choice test items. Each of the expected trends appeared in the data; however, only the latter one, the Training Method × Test Mode interaction, was statistically significant in the overall analysis of variance. However, one-tailed t tests indicated that the program led to significantly greater achievement than the summary among high school students ($t = 2.14$, $df \simeq 90$, $p < .05$) but not among college students ($t = 1.23$, $df \simeq 90$, $p > .05$); and on the delayed test ($t = 2.15$, $df \simeq 90$, $p < .05$) but not on the immediate test ($t = 1.22$, df 90, $p > .05$).

The fact that receiving the immediate achievement test produced a significant increment on the delayed test is not surprising, since it has been well documented that responding to questions or test items during or shortly after training facilitates later performance, even when, as in the present case, no knowledge of results is provided (Michael & Maccoby, 1961; Rothkopf, 1966; Spitzer, 1939). There is the question of whether it is more effective to intersperse questions within the instructional materials, such as is done in a program, or more effective to ask a series of questions after a relatively lengthy presentation. The latter alternative proved more potent in the present study. Considering only performance on the delayed achievement test, it made a small (and nonsignificant) difference whether *S* completed the program (61.3%) or the summary (55.7%) but it made a somewhat larger (and significant) difference whether he received the immediate achievement test (63.5%) or not (53.4%). Among high school *S*s, the program and the immediate test produced increments of the same size and these increments were additive. However, with respect to college undergraduates, for whom the program was designed, the optimum treatment was the summary followed by the immediate test (85.9%). This is an instance in which the teach-and-test policy often denigrated by advocates of programmed instruction worked best.

Test mode interacted significantly with both training method and presence or absence of the immediate achievement test. In each case the short-answer items were more sensitive to the treatment variable than the multiple-choice items. It is possible that short-answer items are more sensitive than comparable multiple-choice items to any treatment difference. However the authors prefer a different interpretation. Discounting the fact that it is often possible to eliminate one or more obviously wrong alternatives when considering a multiple-choice item, the presumption is that the two kinds of items require associative learning in about the same measure. The big difference between the item types is in the requisite level of

TABLE 3
MEAN TRAINING TIME IN MINUTES

Condition	High school	College
Program	74.31	57.02
Summary	14.06	14.72

response learning. An *S* will not be able to emit a poorly integrated response of low strength on a short-answer item, but he may be able to pick the response term from among a set of alternatives. The explanation for the greater sensitivity of the short-answer items in this experiment is that both the program, because of the overt response requirement, and the opportunity to practice the achievement test enhanced response learning. If this line of reasoning is correct, multiple-choice items might be as sensitive as short-answer items to treatments which do not differentially affect response learning.

The results of the experiment reported herein do provide some support for the rationale behind such program features as redundancy and overt responding. Nonetheless, as a practical matter, the most noteworthy finding was that for college students the program did not produce demonstratably better achievement than the summary but took a lot more time. The authors want expressly to disavow any broad generalizations based on this single instance. Programs that are superficially similar may have very different instructional consequences. Indeed, this experiment might have come out differently had later sections of *The Analysis of Behavior* been used. If our analysis is correct, whether a particular program will outperform a summary will depend upon the distance between actual entering behavior and desired terminal behavior, that is, "difficulty;" whether entering behavior is over- or underestimated; whether the programmer has a bias toward "overkill" in the amount of redundancy included; whether empirical techniques are employed in program construction, testing, and development that guard against superfluous redundancy as well as detect gaps in the task analysis. Not mentioned previously, but obviously important, are such additional factors as the completeness of the task analysis and the adequacy of the design of individual frames.

A program with one or more defects may fail to outperform a summary. Most defects cannot be found in a simple examination of a program. Nor is the demonstration that students who complete a program do well on a posttest a guarantee of freedom from defects; students who get some other form of instruction might do better in less time. The authors should like to propose that as a general quality control procedure those who develop programs accept responsibility for demonstrating that their programs outperform summaries. The textbook-style summary of the material in a program makes a feasible trial horse because it can be prepared inexpensively, almost by formula. Because it contains minimal redundancy, a summary would be especially useful in detecting superfluous redundancy, but it could also provide a yardstick to gauge other shortcomings. Finally, if it were the common practice to compare programs with summaries as a step in validation, generalizations about the limits of programming techniques, as they are known today, with various populations and subject matters might thereby arise.

REFERENCES

COULSON, J. E., & SILBERMAN, H. F. Effects of three variables in a teaching machine. *Journal of Educational Psychology*, 1960, **51**, 135–143.

HOLLAND, J. G., & SKINNER, B. F. *The analysis of behavior*. New York: McGraw-Hill, 1961.

MARKLE, S. M. Empirical testing of programs. In P. H. Lange (Ed.), *Programmed instruction*. Part II. Chicago, Ill.: University of Chicago Press, 1967.

MICHAEL, D. N., & MACOBY, N. Factors influencing the effects of student participation on verbal learning from films: Motivating versus practice effects, "feedback," and overt versus covert responding. In A. A. Lumsdaine (Ed.), *Student response in programed instruction*. Washington, D. C.: National Academy of Sciences, National Research Council, 1961.

PRESSEY, S. L., & KINZER, J. R. Auto-elucidation without programing. *Psychology in the Schools*, 1964, **1**, 359–365.

ROTHKOPF, E. Z. Learning from written instructive material: An exploration of the control of inspection behavior by test-like events. *American Educational Research Journal*, 1966, **3**, 241–249.

SPITZER, H. F. Studies in retention. *Journal of Educational Psychology*, 1939, **30**, 641–656.

WILLIAMS, J. P. Comparison of several response modes in a review program. *Journal of Educational Psychology*, 1963, **54**, 253–260.

(Received November 13, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 6, 388–394

THE APPROACH-AVOIDANCE PARADIGM AS A MODEL FOR THE ANALYSIS OF SCHOOL ANXIETY[1]

JAMES A. DUNN
Harvard University

A Dollard and Miller approach-avoidance paradigm was used to predict age, sex, and social-class differences in childrens' school anxiety using as a basis the degree to which *S*s reported liking and valuing the academic aspects of school. Analysis of variance was applied to data collected from 480 public school children. Significant results were obtained in 8 out of 10 analyses. Major findings were: (*a*) as children grow older negative affect for school increases; (*b*) lower-class children report greater school anxiety than middle-class children; and (*c*) lower-class children report more positive affect for the social and the academic aspects of school than middle-class children. Other lesser findings were also reported.

Academic achievement presumably is not only a function of the instructional stimuli which impinge on a learner; but also of the affective state of the learner during the period of that impingement. During the past decade, research on systems theory, programmed instruction, computer simulation, and the like (e.g., Lumsdaine & Glaser, 1960; Ryans, 1963, 1964; Suppes, 1966) has materially advanced one area of knowledge concerning human learning. Somewhat less attention has been devoted to understanding the influence of motivational, that is, affective-attitudinal, factors.

Most conceptualizations of motivation have either been unidimensional, varying only in intensity, as in simple drive theory; or bidimensional, varying in intensity and directionality. Dollard and Miller (1950) extended this latter conceptualization to permit the consideration of multidirectionality.

In the tradition of Lewin (1936), the directionality of drive was a function of the valence characteristic of the goal. Generally, the theoretical problem of multidirectionality was resolved either by positing an alternating valence condition of the object which was due to a changing, often oscillating, need structure of the organism; or by the juxtaposition of objects having different valences, such as food and shock grids.

Regarding human learning, at least two attitudes may operate to motivate learning behavior: (*a*) affect for the material to be learned; and (*b*) the perceived value of the material to be learned. Consequently, some understanding of those factors, especially as regards school-relevant attitude and affect states, would seem to be necessary if one is interested in maximizing children's school performance.

Two sizable modern studies of children's attitudes have been carried out; one by Jersild and Tasch (1949), the other by Witty (1960). Both contained sections dealing with children's school attitudes.

Jersild and Tasch's study was based on questionnaire data collected from over 3,000 children. One of their more striking findings was a "decline, with age, of children's educational morale." At the elementary school level Jersild found only 1% of children's wishes about school to be of a derogatory nature. This figure increased to 10% with junior high school pupils.

Using data from 2,000 pupils in Grades 3–9, Witty found that children liked those subjects best in which they received their best grades. He also found that boys were less interested in the academic aspects of school than girls.

Recently Phillips (1966) has reported

[1] This research was supported, in part, by NH Grant No. 01428 and the Interprofessional Research Commission on Pupil Personnel Services. It was based on a paper presented at the American Psychological Association annual convention, New York City, New York, 1966.

an extensive, semilongitudinal study of school anxiety, and its antecedents, based on a sample of approximately 600 children. One of Phillips' major findings, in addition to significant sex findings, was a relatively high level of school anxiety among minority-group school children, that is, Mexican-American and Afro-American elementary school children.

Theory

The basic assumption underlying this study is that there should be some relationship between the degree to which a child is anxious about school, and the degree to which he likes or dislikes and values or devalues school.

A positive orientation on both of these dimensions, affect and value, would seem to be of crucial importance for a child's adjustment to, and his sustained performance in, school. The pupil who both likes to study academic subjects and who considers the study of academic subjects important would have a more positive attitude toward learning and would, presumably, show less anxiety in his performance than a peer who valued academic achievement, but was negatively attracted (i.e., repelled) by the character of the work involved. The applicability of the approach-avoidance paradigm to the latter instance would seem to be clearly patent. A child who considers academics important but dislikes them would be in an approach-avoidance situation and hence presumably under a good deal of personal stress in precisely those situations calling for a high level and quality of academic performance.

The simplest defense for a child in such a situation would be, of course, either psychological withdrawal via daydreaming and/or rationalization, or physical withdrawal via truancy and/or school dropout. If withdrawal, either physical or psychological, is impossible, one would then expect to find a high degree of anxiety, hostility, and negative affect.

If school anxiety does, in fact, derive from complex attitude patterns toward school, to the extent one would expect to find age, sex, and social-class differences in those attitudes, one could also expect to find age, sex, and social-class differences in the resultant school anxiety.

The present study, then, involves two sets of hypotheses, the first dealing with age, sex, and social-class differences in the basic school attitudes, that is, the affect-value attitudes; the second, a superordinate set, deriving from the former, dealing with predicted age, sex, and social-class differences in school anxiety.

Hypotheses

It was predicted that the adolescent, as contrasted to the preadolescent, would (*a*) place greater value on the academic aspects of school, but because of the increasing stress associated with school and presumably an increasing encounter with negative evaluations of his work, the adolescent would (*b*) have less positive affect for the academic aspects of school.

Because girls in our society typically have less social freedom and less need for vocational skills, and because they typically encounter more academic success in their early school years, it was hypothesized that girls would (*c*) have more positive affect toward the academic aspect of school, but would (*d*) value those aspects less than their male counterparts.

Because of the growing social awareness of the crucial role of education for upward mobility, it was predicted that lower-class children, as contrasted with middle-class children, would (*e*) place more value on the academic aspects of school than middle-class children; but because of their more frequent encounters with school failure, especially in the early years, they would (*f*) like the academic aspects of school less.

On the basis of the Dollard and Miller (1950) approach-avoidance paradigm, it was hypothesized (*g, h, i*) that if males, adolescents, and children of lower socioeconomic status did, in fact, value the academic aspects of school but had relatively little positive affect toward those aspects, then those males, adolescents, and lower socioeconomic class children would manifest more anxiety about school than

would preadolescents, girls, and upper-middle-class children.

Although the basic intent of the present study was to investigate these hypotheses, because of the opportunity afforded, an ad hoc investigation of children's attitudes toward the social aspects of school was also included. In general, it was tentatively expected that adolescents would no longer consider the social aspects of school either as important or as positive as would preadolescents; that, because of their greater dependence on school for social contacts, adolescent girls would both value and like the social aspect of school more than boys; and that lower-class children would neither like nor value the relatively rigid, middle-class supervised, social aspects of school as much as middle-class children.

METHOD

Because of the desirability of collecting the data in situ so that familiar school and classroom cues would be present for the child as he responded to the instrument, the classroom was chosen as the unit of data collection. The child, however, was the unit of data analysis.

Subjects

Data were collected from 30 classrooms, five classes from each social class at each of three grade levels (see Table 1). Forty *S*s were then drawn, using a table of random numbers, for each

TABLE 1

CHARACTERISTICS OF TOTAL SAMPLE

Socioeconomic class	Pupil distribution		
	Male	Female	Total
Grade 5			
Middle	78	58	136
Lower	65	67	132
Total	143	125	268
Grade 7			
Middle	53	59	112
Lower	58	48	106
Total	111	107	218
Grade 9			
Middle	56	49	105
Lower	52	65	117
Total	108	114	222
All grades	362	346	708

Note.—A total of 30 classrooms were sampled, 5 for each socioeconomic class in each grade.

TABLE 2

AGE OF SUBJECTS SELECTED FOR ANALYSIS: MEANS AND STANDARD DEVIATIONS BY GRADE, SOCIOECONOMIC CLASS, AND SEX

Socio-economic class	Fifth grade		Seventh grade		Ninth grade	
	Male	Female	Male	Female	Male	Female
Middle						
$\bar{X}$	10.23	10.13	12.00	12.15	14.13	14.00
SD	.42	.46	.23	.43	.72	.51
Lower						
$\bar{X}$	11.03	10.40	12.88	12.75	14.78	14.18
SD	.83	.63	.88	.74	.80	.64

of the 12 cells of the analysis. Thus, the subsample on which data analysis was executed was 480 students.

Table 2 summarizes the age by sex characteristics of the data-analysis sample.

The upper-middle-class portion of the sample was drawn from a suburban school system (from a city coded OP) where the median years of schooling completed by persons 25 years of age or older was 12.4. Less than 3% of the potential male labor force was unemployed. The median family income was $8,657.

The lower socioeconomic portion of the sample was drawn from the inner city area of a large metropolitan city (coded D) where the median years of schooling completed was slightly less than 9. Over 20% of the male population was unemployed, and the median family income slightly less than $3,500. At least half of the families in this latter group could be defined as poverty families by Office of Economic Opportunity standards.[2]

Instrumentation

The *S*s were asked to rate on a 6-point scale the degree to which they valued, and later, the degree to which they liked, eight different school activities ranging from learning about science and nature to playing games or sports. Four ratings pertained to the academic aspects of school, four pertained to the social aspects of school. The rating scales were part of a larger questionnaire given in two sections, 1 week apart (see Morse, Bloom, & Dunn, 1961). Anxiety was measured with a modified version of Sarason's Test Anxiety Scale, a scale which factor analysis has indicated to be a more appropriate measure of school anxiety than of simple test anxiety per se (Dunn, 1964).

Data Analysis

A three-factor analysis of variance (Winer, 1962) was used to determine the main and interaction effects of age, sex, and social class. Filter means analysis, a partitioning program developed for use by the University of Michigan Institute for

[2] This information was derived from the 1960 United States Census data.

Social Research, was then used to determine the precise nature of the interaction effects that proved to be significant. Because it iterates through all orthogonal combinations, filter means analysis may be used to isolate the source of interaction effects previously found to be significant through traditional analysis of variance.

Results

Analysis of variance results are summarized in Table 3; eight of the nine analyses of variance yielded significant F ratios. Figures 1–3 summarize some of the more interesting data trends.

Lower-class children indicate a higher level of school anxiety than middle-class children (significant at .05 level). This result was due almost entirely to the responses of children at the elementary school level, however (see Figure 1). The differences in anxiety between adolescents and preadolescents was not significant. An inspection of the group means suggests a trend in the predicted direction, however. With regard to the predicted sex differences, the converse was found. Girls were found to have a significantly higher (.05) level of school anxiety than boys. Filter means analysis indicated this ANOVA main effect to be due, in large measure, to differences at the seventh grade level. There were, in fact, counter-indications at the fifth grade level.

Concerning affect-value patterns, it was found that as students grow older, they like all aspects of school, social as well as academic, less and less (significant at the .01 level). This was true for the most part regardless of sex or social class. The rate of decrease in affect for the academic, as well as the social, aspects of school was less for lower-class children, however. Also, as upper-middle-class children grow older, they tend to devalue, as well as dislike the academic and social aspects of school (significant at the .01 level). This was not the case for lower-class children.

TABLE 3

Summary of Hypotheses and Results

Variables	Hypotheses	Level of significance	Comment
Age	Adolescents will:		
	dislike academics	.01[a]	—
	value academics	.01[b]	Middle-class adolescents showed a significant decrease in value for academics.
Sex	Girls will:		
	like academics	.05[b]	Preadolescent girls liked academics more than boys.
	value academics	.01[a]	Girls claimed they valued academics more than boys; preadolescent girls claimed they valued academics more than any other group.
Social class	Lower-class children will:		
	dislike academics	.01[a]	Lower-class children report greater positive affect for academics than the middle-class children.
	value academics	.01[b]	Confirmed for lower-class adolescents only.
Anxiety	High anxiety level for:		
	adolescents	*ns*	—
	males	.05[b]	Significant difference at the seventh-grade level only, where males have lower anxiety scores than females.
	lower-class children	.05[a]	—

[a] ANOVA main effect.
[b] ANOVA interaction effect.

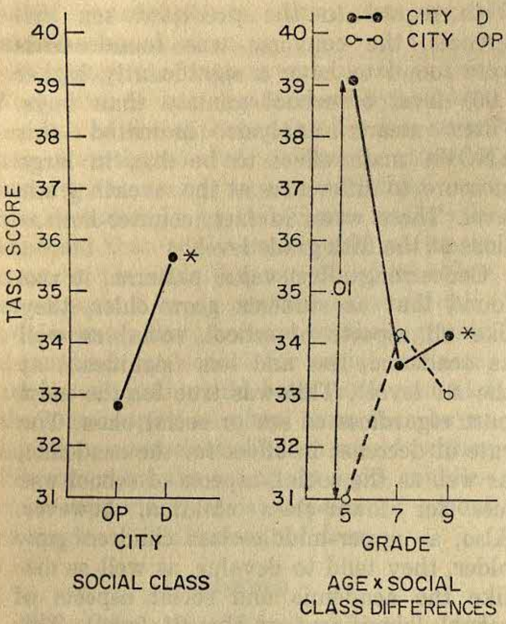


FIG. 1. School anxiety levels. (Note, in Figures 1–3, * signifies ANOVA differences significant at .05 level; ** signifies significant at .01 level; values between vertical arrows signify levels of significance of filter means differences.)

There was no decrease in their reported level of perceived value for the various aspects of school.

Regarding sex differences, elementary school girls reported that they liked the academic aspects of school more than elementary school boys (.05 level), but there was a steady decline with age in both (*a*) the degree to which girls liked the academic aspects of school (.01 level) and (*b*) the degree to which they valued them (.01 level). Girls, however, tended to remain higher than boys in the value they placed on academics. Interestingly, there were no sex differences in the degree to which boys or girls in this sample valued or enjoyed the social aspects of school. There was a decrease with age, though, in the degree to which girls reported liking the social aspects of school (.01 level).

Regarding lower-class differences, lower-class adolescents reported they both valued and enjoyed the academic aspects of school more than upper-middle-class children (.01). This was true for all grade levels and for both sexes. They also claimed to value the social contacts more (.01).

There were no social-class differences in the degree to which children enjoyed the social aspects of school, however. Judging from these results, it is possible that the social aspects of school are largely independent of the value structure of school authorities and are, in fact, far more in the hands of the peer group than in the adult authorities.

DISCUSSION

In view of the author's contention that human motivation, as far as school behavior is concerned, should be given multidimensional consideration, and that an individual's motivational state at any point in time probably involves a complex hierarchy of a good many approach-avoidance values, the success that was achieved with a simple two-value paradigm was surprising. Success was mixed though. It is probably the case that, as far as adolescents are concerned, there are many other school factors far more salient for the determination of school approach-avoidance than those chosen for the present study. When the question is polarized

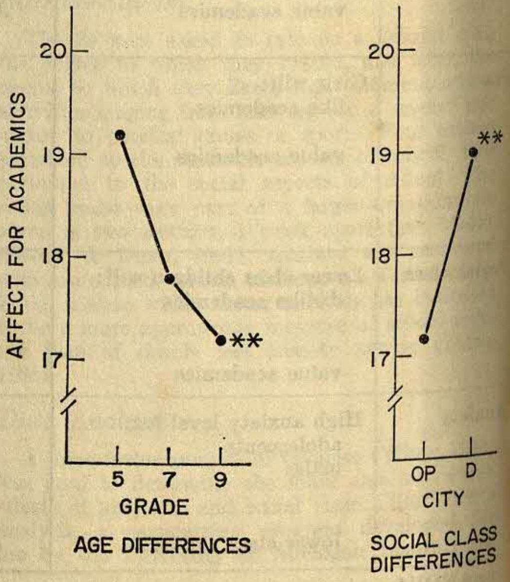


FIG. 2. Affect for academics.

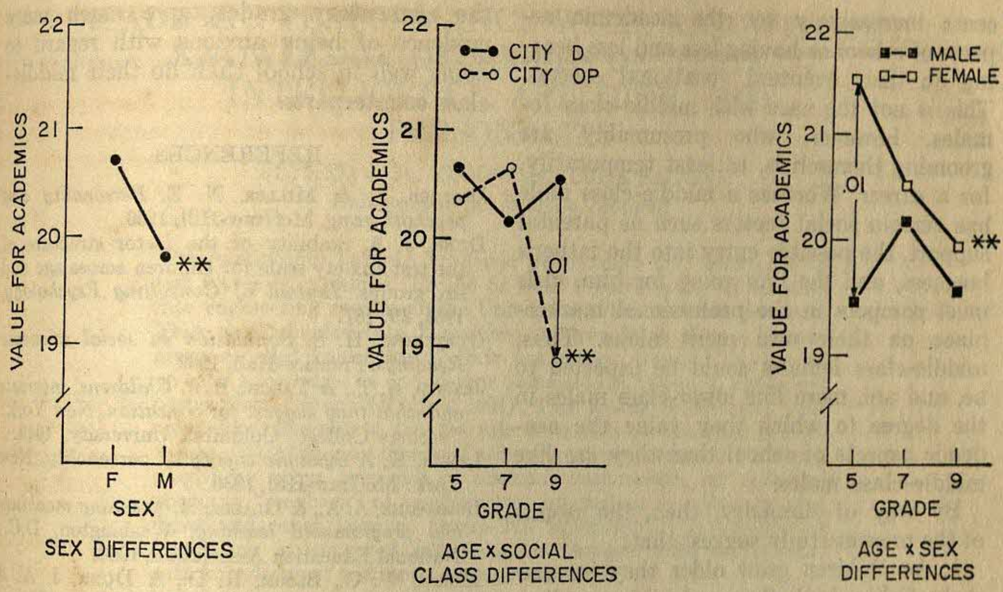


FIG. 3. Value for academics.

around the issue of social-class differences, however, the affect-value dimensions seem to take on much greater saliency; and hence are much more intimately associated with school anxiety.

The age and sex findings of the present study, with regard to attitudes toward school are by and large consonant with earlier findings. It is the social-class findings of the present study that are of special interest. Especially in view of the fact that they challenge some of the stereotyped notions that have long been entertained with respect to lower-class children in middle-class school settings. It appears, for example, that the lower-class child both appreciates and values the academic aspects of school much more than he has been given credit for in the past. In addition, he also is apparently very much concerned with doing well, at least as far as this is suggested by school anxiety.

These last findings, with regard to social class and anxiety, have since been corroborated by Sheila Feld[3] at the National Institute of Mental Health. Dr. Feld has also found lower-class children to have a higher degree of test anxiety than middle-class children.[4]

Two separate hypotheses regarding why this should be so may be suggested. One holds that the lower-class child's school anxiety is, in fact, reality oriented inasmuch as he typically has met with a high degree of failure in school activities, hence confrontation with further possible failure is anxiety arousing. The other explanation holds that for the lower-class child educational success is a necessary requisite for upward mobility, thus more of his future is at stake in school and in testing situations than is the case with the middle-class child.

If only school anxiety scores are inspected, it would appear that the former hypothesis has the edge. But if value for academics is also considered, one is met with the peculiar pattern of lower-class values for academic pursuits remaining reasonably high whereas middle-class values for academics fall off drastically at adolescence, and especially for males.

It is possible that middle-class adoles-

[3] Personal communication, 1966.

[4] Phillips, in a study reported after the present study was completed, found similar results.

cents increasingly see the academic aspects of school as having less and less bearing on their eventual vocational success. This is not the case with middle-class females, however, who presumably are grooming themselves, at least temporarily, for a career. Whereas a middle-class male has certain social factors such as parental support, the possible entry into the father's business, and the like going for him, girls must compete in the professional marketplace on their own merit alone. Thus, middle-class females could be expected to be, and are, more like lower-class males in the degree to which they value the academic aspects of school than they are like middle-class males.

By way of summary, then, the results of the present study suggest that:

1. As children grow older they increasingly dislike both the academic as well as the social aspects of school.

2. In elementary school, girls like and value academics more than boys, but these sex differences disappear as children grow more and more to dislike and devalue the academic aspects of school.

3. Lower socioeconomic children at all ages and both sexes report liking the academic aspects of school more than upper-class children. As they grow older and move into adolescence, lower socioeconomic class children continue to value the academic aspects of school whereas their upper-middle-class counterparts come to increasingly dislike and devalue them.

4. As lower-class children grow older they report that they also value the social aspects of school more than upper-middle-class children.

5. Lower-class children, especially in the elementary grades, give much more evidence of being anxious with regard to doing well in school than do their middle-class counterparts.

REFERENCES

DOLLARD, J., & MILLER, N. E. *Personality and psychotherapy*. McGraw-Hill, 1950.

DUNN, J. A. Stability of the factor structure of the test anxiety scale for children across age and sex groups. *Journal of Consulting Psychology*, 1964, **29**, 187.

GUETZKOW, H. S. *Simulation in social sciences: Readings*. Prentice-Hall, 1962.

JERSILD, A. T., & TASCH, R. J. *Childrens' interests and what they suggest for education*. New York: Teachers College, Columbia University, 1949.

LEWIN, K. *A dynamic theory of personality*. New York: McGraw-Hill, 1936.

LUMSDAINE, A. A., & GLASER, R. *Teaching machines and programmed learning*. Washington, D.C.: National Education Association, 1960.

MORSE, W. C., BLOOM, R. D., & DUNN, J. A. *A study of school classroom behavior from diverse evaluative frameworks: Developmental, mental health, substantive learning, group process*. Ann Arbor: University of Michigan, School of Education, 1961.

PHILLIPS, B. N. An analysis of causes of anxiety among children in school. United States Office of Education Final Report, Project No. 2616, Austin: University of Texas, 1966. (Mimeo)

RYANS, D. G. An information systems approach to education. Systems Development Corporation, Technical Memorandum 1495, Santa Monica, 1963.

RYANS, D. G. Systems analysis in educational planning. Systems Development Corporation Technical Memorandum 1968, Santa Monica, 1964.

SUPPES, P. The uses of computers in education. *Scientific American*, 1966, **215**, 207–220.

WINER, B. J. *Statistical principles in experimental design*. McGraw-Hill, 1962.

WITTY, P. A. *A study of the interests of children and youth*. Washington: Government Printing Office, 1960.

(Received January 16, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 6, 395–401

ACADEMIC PERFORMANCE WITH, AND WITHOUT, KNOWLEDGE OF SCORES ON TESTS OF INTELLIGENCE, APTITUDE, AND PERSONALITY[1]

ALFRED J. M. FLOOK[2] AND USHA SAGGAR[3]
University of St. Andrews, Scotland

The relationship between academic performance and knowledge of test scores is examined. The 60 *S*s were 1 year's intake of students into engineering courses at the University of St. Andrews, Scotland. They were divided into 2 matched groups: *S*s in Group K were given detailed knowledge of their test scores; Group NK received no such knowledge. In end-of-year examinations, Group K performed better than Group NK ($p < .001$). This finding is discussed with particular reference to (*a*) the role of anxiety in academic performance, (*b*) Atkinson's theory of achievement motivation, (*c*) the part played by other correlates of academic performance, and (*d*) the optimum "feedback" of psychometric information to university students and other learners. Group K's superiority is interpreted as originating in improved self-evaluation through social comparison, with knowledge of test scores acting catalytically.

It is sometimes suggested that examination results suffer if students are at some earlier stage informed of their scores on tests of ability, aptitude, or personality. According to this argument, the low scorers become demoralized and the high scorers too complacent, so that the later academic performance of both is inferior to what it would have been had they been kept in ignorance of their test scores. In the United Kingdom and elsewhere this is a common objection to the disclosure of test scores to students. In support of it, anxiety level is usually postulated as being the key variable intervening between knowledge of test scores and academic performance. Against the background of studies (e.g., Lynn & Gordon, 1961; Savage, 1962) suggesting that the relationship between anxiety and performance obeys the curvilinear Yerkes-Dodson principle, it is held that students who know their test scores are low are impeded by overanxiety in their later work; whereas those who know their scores are high, in their self-satisfaction, fall below that "happy medium" level of anxiety that spurs them on to their best performance.

Though plausible, the argument is open to the objection that the crucial factor is the form in which the feedback information is given. A mild version of this counterclaim asserts merely that "good" presentation cancels out any ill effects, while a stronger version claims that the feedback may take a form that outweighs any ill effects, thereby producing a net gain.

"To tell or not to tell" is a question of both practical and theoretical importance, yet reported attempts to obtain an experimental answer are lacking. The present exploratory study was included in a larger investigation (Saggar, 1961) in an attempt to resolve the question experimentally.

METHOD

Hypotheses and Experimental Design

The broad research hypothesis to be tested is that knowledge of scores on tests of ability, aptitude, and personality results in poorer academic performance. The *S*s were divided into two groups,

[1] The authors thank the Faculty of Applied Science in the University of St. Andrews for its unstinted assistance and tolerance throughout the 2 years in which data were being collected. Also, they would like to express their special gratitude to the 154 students who gave up a considerable amount of their time, most of it to endure intensive testing and interrogation, in order to reveal enough of their private and academic lives to make the investigation possible.

[2] Now at the University of Dundee, Scotland.
[3] Now at the Department of Clinical Psychology, Royal Dundee Liff Hospital, Dundee, Scotland.

between which the differences on all the test variables were insignificant ($p > .20$).

The two subhypotheses are that knowledge of test scores lowers the subsequent academic performances of students with (*a*) high scores and (*b*) low scores on the tests. The two groups were formed in a way which permitted the subdivision of each into three comparable sections, made up of matched pairs of *S*s with scores in the top, middle, and bottom thirds of the distribution of intelligence-test scores.

Subjects

The *S*s were 60 students (1 female and 59 male) admitted to engineering courses at the University of St. Andrews in October 1960. Of the complete intake of 62 students, only 2 declined to take part in the investigation. One of the 60 volunteers did not complete the degree examinations in June 1961, which were used as the criterion in this experiment. Consequently, he and his pair in the other group were both omitted from the analysis of the results. Except where indicated, therefore, the results given and discussed in this paper relate to 58 *S*s in two matched groups. However, the experiment formed part of a broader inquiry covering the intakes of the previous 2 years as well, and in the present paper reference is made explicitly to the collective data for the three consecutive intakes, involving 154 students, where further insight seems to be obtainable from them.

Materials

The tests administered were the AH5 Group Test of High-Grade Intelligence (AH5; Heim, undated), the Engineering and Physical Science Aptitude Test (EPSAT; Moore, Lapp, & Griffin, 1943), and the Maudsley Personality Inventory (MPI; Eysenck, 1959). Two specially designed questionnaires were also used, one concerned with methods of study, and the other with biographical data about each *S*, his home environment, motivation, worries, etc. Additional personal information was obtained in a series of individual interviews.

Procedure

The general situation regarding the collection of data was explained beforehand to each *S* in a personal note from the Dean of the Faculty of Applied Science. In this he made it clear that the Faculty was sponsoring the investigation, that participation in it was voluntary, and that information obtained about *S*s at all stages of the inquiry would be treated in strict confidence.

The data were collected in the following stages:

Administration of tests and biographical questionnaire. This was done at the start of the first term in October 1960. When all the tests had been scored, *S*s were divided into two matched groups of 30 students. These then passed through the stages described below, except for the one *S* who did not complete the degree examinations.

First basic interview. Early in the second term, all *S*s were given an individual interview, during which they answered questions on methods of study. At the start of their interview, those *S*s in Group K were given their own scores on each of the three tests, together with enough normative data for them to be able to see how they stood in relation both to the original standardization groups and to their own classmates (i.e., the group of 60 entrant students). A special effort was made to present this information in a way that would be encouraging rather than intimidating, without being misleading. It was pointed out that the association between ability and academic success, while positive, is not overwhelmingly strong in a highly selected university population, so that at that level other factors (strong motivation, good study methods, hard work, etc.) assume a special importance. The other *S*s, Group NK, were told that no knowledge of their test results could be given to them for the time being, as that variable was being controlled in order that its effect on performance could be determined.

Supplementary interview on methods of study. Later in the second term, an interview was given to any *S* of either group who had asked for such advice in the first interview.

Second basic interview. Early in the third term, all *S*s were questioned on their performance in the class examinations at the end of the second term, and various data necessary for the main part of the investigation were collected.

A striking feature of all the interviews was the great willingness of *S*s to discuss their personal problems. It seems clear that the junior author, who conducted the interviews, had gained *S*s' confidence in the manner of an effective student counselor.

First-year degree examinations. In June 1961, all *S*s except one took degree examinations in mathematics, physics, and chemistry. The marks for these three examinations were normalized with a mean of 55 and a standard deviation of 10, and then added for each student to form a criterion score. This composite mark is used in Table 1 to compare the examination performance of the two groups.

RESULTS

Main Hypothesis

The main hypothesis is that knowledge of test scores does result in poorer academic performance. Its opposite is a compound of two alternative counterclaims: that knowledge of test scores makes no difference in academic performance (the null hypothesis); or that it results in better performance. Consequently, there were, in effect, two research hypotheses available, predicting deviation from the null hypothe-

sis in opposite directions. To decide between them, the null hypothesis was tested by the two-tailed version of statistical tests appropriate to the matched-pairs experimental design.

The outcome, detailed in Table 1, is a highly significant difference in favor of Group K, its probability being either a little below or above the .001 level, according to the test applied.

Subhypotheses

The subhypotheses are that knowledge of test scores lowers the subsequent academic performance of high-scoring and low-scoring *S*s. The results, detailed in the lower rows of Table 1, were:

Subjects with high test scores. The difference in performance of *S*s with high test scores was not significant, according to the two-tailed tests used. However, it favored Group K, not Group NK, and it almost reached the .05 level of significance for the one-tailed versions of the tests.

Subjects with low test scores. The same trend was unmistakable for the 10 low-scoring pairs. Here the 10 *S*s who had been told their test scores proved decisively superior to their counterparts in Group NK, the difference easily exceeding the .01 level of significance and nearly reaching the .001 level.

TABLE 1

GROUPS K AND NK (AND THEIR SECTIONS) COMPARED ON CRITERION SCORES

Subjects	Group K		Group NK		Comparisons	
	M	*SD*	*M*	*SD*	T[a]	t[b]
Whole groups (N = 29)	178	19.1	156	24.1	61*[d]	3.89**
Sections[c]						
Top (N = 10)	182	19.0	160	32.2	13	1.72
Middle (N = 9)	176	21.6	158	19.4	9	1.60
Bottom (N = 10)	176	16.4	149	15.8	0*	4.47*[e]

[a] Two-tailed Wilcoxon matched-pairs signed-ranks test.

[b] Two-tailed t test for paired samples.

[c] Of the two groups in terms of AH5 scores, which were 54–42, 41–36, and 36–23, respectively.

[d] $p = .0014$

[e] With 9 *df*, a two-tailed p of 0.001 corresponds to a t of 4.781.

* $p < .01$.

** $p < .001$.

There are two general features that deserve comment. First, although only two of the four differences reported in Table 1 reached significance, all four were in the direction of Group K's superiority to Group NK. Second, the bottom section of Group K had a mean criterion score equal to that of the middle section of Group K, and larger than that of every section of Group NK. In other words, there is an indication that, while knowledge of test scores conferred an advantage throughout the range of ability in Group K, its impact was greatest on the bottom third, with the lowest test scores. (The possible connection between this and the fact that only 2 years previously the first-year failure rate had climbed to 33% is one of the questions considered.)

DISCUSSION

Considering how few *S*s were available, the result is remarkably clear-cut. This highly significant difference must have come entirely from an above-normal performance by Group K, if Group NK can justifiably be regarded as a control group. That seems a reasonable view, as test scores are not normally available to these students. However, the difference could conceivably have arisen entirely from a below-normal performance by Group NK, or it might have been a joint effect in Group K's favor which emerged because the performance of both groups had deviated, up or down, from the norm of previous years. To clarify the point, the failure rate for 1956–60 was analyzed. The information available was not such as to allow an absolutely definitive conclusion to be drawn, but the analysis did suggest very strongly that Group K had performed above the recent norm, as well as significantly better than Group NK, as a result of having been told their test scores.

Since such an outcome is so satisfactory educationally, and as test scores are normally available nowadays for so many students and other "learners," the question of the generality to be granted to this re-

sult has special practical importance. The *S*s came from a narrow band of the educational spectrum. Not only were they all university students, drawn from one university, but they were all studying engineering, and all but one were male. However, they may reasonably be regarded as being a representative sample of that special group, for they comprised all but 2 of a year's intake of 62. Furthermore, there is no obvious reason why the result should be unique to such students, but the extent to which it can be generalized beyond them is a question needing further study.

Another striking feature of the result is its unexpectedness. Not only does it run counter to the opinion prevalent in Great Britain that knowledge of test scores depresses subsequent academic performance, but also it goes beyond the intermediate view that this factor is immaterial, by creating the strong presumption that it has a facilitating influence. Such a "booster" effect, if confirmed in replications, would have theoretical and practical implications that make it important to consider possible explanations.

The Role of Anxiety in Academic Performance

Previous research suggests a curvilinear relationship[4] between anxiety and performance, with differential anxiety as the mediating factor. Even though no direct measure of anxiety was available for *S*s of this experiment, it is obvious that its results refute a simple differential anxiety hypothesis. Evidently, some more complex role for anxiety as an intervening variable is needed if it is to fit the facts—a point stressed by other workers in this field (e.g., Stein, 1963).

Atkinson's Theory of Achievement Motivation

One of the purposes of Atkinson's (1957) theory is to account for performance level when only one task is presented to *S*s in a competitive setting. That was the situation confronting Groups K and NK, whose members were taking a compulsory curriculum. Most of them may be assumed, in view of the long record of academic success necessary to gain admission to a British University, to have been stronger in the motive to achieve success (M_S) than in the motive to avoid failure (M_{AF}). And Atkinson, in applying his theory to persons in whom $M_S > M_{AF}$, concludes that the strength of motivation to perform a task (when no alternatives are offered) should be greatest when such persons are most uncertain about the result, that is, when the probability of their success (P_s) is .50. And so, performance level, through which strength of motivation is expressed, should be a bell-shaped function of P_s (as shown in Figure 1) when differences in ability are controlled, as in the matched-pairs design used for this experiment.

There is evidence (Atkinson, 1964; Atkinson & Feather, 1966) that individual differences in intelligence or aptitude may serve as cues to define a person's P_s in a competitive academic setting. Conceivably, the members of Group K, having a clear indication of their relative standing on the tests, developed sharply defined P_s values scattered widely on either side of the critical central value. By contrast, the mem-

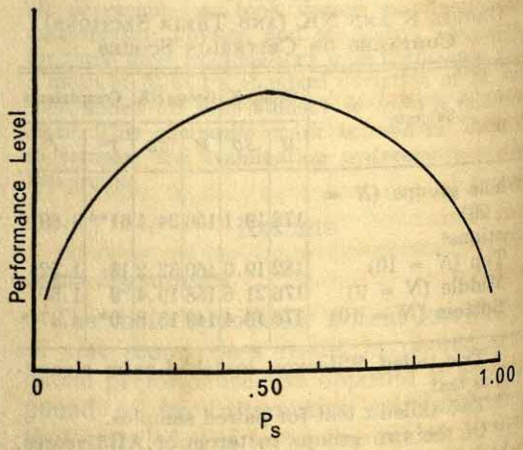


FIG. 1. Strength of motivation to achieve, or to avoid failure, as a function of subjective probability of success. (A slightly modified form of Figure 1 in Atkinson, "Motivational Determinants of Risk-Taking Behavior," *Psychological Review*, Vol. 64, 1967, 359–372).

[4] In the form of an inverted-U curve, like that shown in Figure 1, if anxiety is represented by the abscissa.

bers of Group NK were all alike in their ignorance of their test scores, and that greater uncertainty could have had the centripetal effect of inclining them to cluster closer to a P_s of .50. Under such conditions, the theory would lead one to expect that Group NK would perform better than Group K. However, the information available on the motivational dynamics of the experimental situation is ambiguous enough for the theory to yield, on other assumptions, a prediction in favor of Group K—in line with the actual result. The possibility of such contradictory speculation highlights the essential point that the clear need now is for a series of experiments in which the crucial intervening variables in a complex situation are identified, measured, and controlled in the kind of systematic and penetrating analysis which the present exploratory study, made under field conditions, was not intended to provide.

The Part Played by Other Correlates of Academic Performance

As already indicated, the experiment formed part of a broader inquiry into the relationship between academic performance and many factors presumed to be influential independent variables. Some of these were, in fact, found to be correlated with academic performance; and it is conceivable that Group K's superiority came from a membership more favorably endowed than Group NK's with these attributes, exerting an influence either antecedent to, or concurrent with, knowledge of test scores.

Five variables found to be associated significantly (Numbers 1, 2, and 3 at the .01 level) with academic success were:

1. The *S*'s report that his work was not interfered with by daydreaming.
2. Normal home background (operationally defined as one which, according to *S*, contained none of the domestic abnormalities on the list used in this inquiry).
3. The *S*'s report of freedom from financial worries.
4. Parent's income being in the top, rather than the middle, section of a specified 3-part range.
5. Parent's occupation falling into Grades I and II, rather than III–VII, of the Hall-Jones scale (Hall & Jones, 1950).

Of these, the first alone may have given some advantage to Group K, whereas the other four tended to favor Group NK. It seems therefore that, at the very least, the matching of the two groups put Group K in no stronger position than Group NK in terms of these five variables, viewed as a whole.

There remains the possibility of explaining Group K's superiority in terms of variables intervening between *S*s' being told their test scores and the degree examinations nearly 6 months later. The first indication of such a chain of events is to be seen in the unequal advantage taken by *S*s of the availability of advice on methods of study. The striking difference is that 24 of Group K, as against 11 from Group NK, asked for the help in this direction which had been offered impartially to all 58 *S*s.

Further light is shed by *S*s' answers to the question, "Approximately how many hours do you devote daily to your studies, that is, apart from the university hours?" Some gave their answer there and then, but others required time to work it out. The latter reported theirs 2 or 3 weeks later. All the data are summarized in Table 2.

TABLE 2

Groups K and NK Compared on Hours Spent Weekly on Private Study, According to Students' Own Replies to Questions in Interview

Reply	Group K		Group NK		Comparisons
	M	*SD*	*M*	*SD*	t^a
Given					
Immediately	15.2^b	3.31	14.5^c	3.35	.51
Later	24.4^d	3.95	13.0^e	3.86	8.14*
Total	20.2	5.79	13.6	3.76	5.08*

[a] Two-tailed test.
[b] $N = 13$.
[c] $N = 12$.
[d] $N = 16$.
[e] $N = 17$.
* $p < .001$.

These figures raise several questions, but the most plausible interpretation seems to be that knowledge of test scores had a delayed-action effect on work habits, inducing the 16 members of Group K who replied later to increase their working week in the light of their sober reflections on the academic future painted for them by their test scores. If so, a similar reassessment may have been made subsequently by their 13 fellow members who had replied on the spot, which would, of course, have swollen the overall difference between the two groups in hours of study.

All this suggests that the simple explanation for Group K's superiority is that collectively they worked harder and more effectively than Group NK, but that is an unenlightening tautology. Greater justice would be done to all the evidence available by suggesting that knowledge of his test scores gave each member of Group K a clearer and more realistic picture of his academic possibilities than that possessed by his counterpart in Group NK, and that this sensitizing experience moved him to take better advantage of the general context of help that was equally available to both groups. In other words, the impetus toward superior performance may have been supplied by improved self-evaluation through social comparison, which has been proposed as a major motive underlying social behavior (Latané, 1966). The fact that a number of Group K's members asked spontaneously for the norms for their own classmates immediately after being given the original standardization norms lends empirical support to this suggestion.

Clearly, once this psychometric information had been given, advice on study methods was a key variable, but it was only one ingredient in a complex of supporting factors. Another important component was the almost continuous accessibility of the junior author to *S*s for the discussion of their work and problems. Being Indian and a woman may have helped her to establish close rapport, which itself might suggest that the results of the experiment reproduce the "Hawthorne" effect. That seems doubtful, however, since the same interest was taken in both groups throughout, the only difference being the quantitatively minute one involving the disclosure of test scores.

The care taken to give Group K their test scores in a positively energizing manner was also intended to be a facilitating ingredient. In this respect, it is notable that it was the bottom third of Group K (in terms of AH5 scores) who contributed most to the overall superiority of their group, and the first-year failure rate (a vital fact of academic life which most entrant students get to know before, or soon after, admission) had reached 33% only 2 years previously. The *S*s in the bottom section of Group K were, therefore, in possession of full information from which they could conclude that they were in the academic danger zone.

In summary, the interpretation suggested is that knowledge of test scores acted like a catalyst, setting up in Group K's members an elaborate chain reaction, culminating in superior academic performance. It did so by first clarifying their relative standing in academic potential, thereby creating an informed concern about future performance that impelled them both to work harder and to exploit fully the facilities available to all for maximizing academic success. This assigns a central activating role to one aspect of anxiety (the phrase "informed concern" being intended to convey that Group K's alertedness took a predominantly facilitating form), but this is speculation, as the experiment included no measure of anxiety specific to academic performance.

Further Research Needed

Several questions have been brought up that point to a need for the experiment to be replicated and systematically extended. Its theoretical implications call for a penetrating exploration of the complex social context of academic performance to elucidate the contributions of psychometric information, of anxiety, and of their interplay with each other and with other variables in the educational setting. Analysis along similar lines would be necessary to clarify the relationship of this experimental

finding to Atkinson's theory of achievement motivation.

As for its practical significance, the finding suggests that the wealth of psychometric data on record about students and other learners in so many parts of the world may have a productive potential that has been overshadowed by the diagnostic and prognostic functions that are now established traditions. Such a claim would rightly provoke a demand for cautious application, in view of the issues still unresolved. One crucial question is: What forms should psychometric "feedback" take so as to produce the best effects? Others are facets of the problem of generality. If confirmed on similar populations, how far could this finding be generalized? Is it specific to certain age or ability levels, to certain subjects of study, to males rather than females? In view, too, of the evidence (McClelland, 1961) for cultural differences in achievement motivation, is it a phenomenon peculiar to certain societies and absent from others? It is clear that the finding has implications, both practical and theoretical, calling for a comprehensive understanding of the social complex in which the effect was embedded.

REFERENCES

ATKINSON, J. W. Motivational determinants of risk-taking behavior. *Psychological Review*, 1957, **64**, 359–372.

ATKINSON, J. W. *An introduction to motivation*. Princeton: Van Nostrand, 1964.

ATKINSON, J. W., & FEATHER, N. T. (Eds.), *A theory of achievement motivation*. New York: Wiley, 1966.

EYSENCK, H. J. *Maudsley Personality Inventory*. London: University of London Press, 1959.

HALL, J., & JONES, D. C. Social grading of occupations. *British Journal of Sociology*, 1950, **1**, 31–55.

HEIM, A. W. *AH5 Group Test of High-Grade Intelligence*. London: National Foundation for Educational Research in England and Wales (undated).

LATANÉ, B. (Ed.), *Studies in social comparison*. New York: Academic Press, 1966.

LYNN, R., & GORDON, I. E. The relation of neuroticism and extraversion to intelligence and educational attainment. *British Journal of Educational Psychology*, 1961, **31**, 194–203.

McCLELLAND, D. C. *The achieving society*. Princeton: Van Nostrand, 1961.

MOORE, B. V., LAPP, C. J., & GRIFFIN, C. H. *Engineering and Physical Science Aptitude Test*. New York: The Psychological Corporation, 1943.

SAGGAR, U. *Psychological aspects of selection for engineering education*. Unpublished Bachelor of Philosophy thesis, University of St. Andrews, 1961.

SAVAGE, R. D. Personality factors and academic performance. *British Journal of Educational Psychology*, 1962, **32**, 251–3.

STEIN, M. I. *Personality measures in admissions*. New York: College Entrance Examination Board, 1963.

(Received September 7, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 6, 402–403

MENTAL RETARDATION, MENTAL AGE, AND LEARNING RATE

ARTHUR R. JENSEN AND WILLIAM D. ROHWER, JR.
University of California, Berkeley

Zigler's hypothesis that mental age (MA) and not IQ determines the rate of learning is examined in the light of empirical evidence comparing the learning rates of normal and retarded children and young adults matched for MA. The results show that learning rate is a function of IQ as well as of MA. In general, children of average IQ learned serial and paired-associate lists significantly faster than retarded young adults with IQs between 50 and 60 but with approximately the same MA as the children. An interaction between IQ, learning rate, and socioeconomic status is also noted.

Zigler has now stated (1967a) and restated (1967b) a central theme of his theoretical position regarding mental retardation that "...it is the MA [mental age] (level) and not the IQ (the relationship of MA to chronological age) that determines the exact nature, including the rate, of learning any task [1967b, p. 579]." Thus, two persons of different chronological age (CA) and different IQ but matched on MA should show similar learning rates.

Weir (1967) has challenged Zigler's statement on essentially the following basis: If MA is a measure of the knowledge an individual has accumulated by a given CA, the *rate* of acquisition of this knowledge is represented by the IQ, which is (MA/CA) × 100. Therefore, contrary to Zigler's position, persons of the same MA but differing in IQ should show different rates of learning, even in short-term learning tasks. There is evidence that Weir's prediction is indeed borne out in the case of laboratory learning tasks.

The obscurities in the argument between Zigler and Weir can be overcome by making a conceptually clear-cut distinction between developmental rate and learning rate. There is much evidence (White, 1965) that mental abilities have a hierarchical structure, the development of which follows a chronological sequence; the milestones of this developmental sequence are marked by the increasing complexity of the cognitive structures (e.g., heuristics, symbolic mediators, strategies, information processing skills) which the individual can bring to bear on solving problems. The ages at which individuals attain these stages of cognitive development are regarded as indexes of developmental rate. But two individuals who are at the same developmental stage and who have arrived at this stage at either the same or at different rates of development, may still differ in the rates at which they can acquire new information. This is distinguished as learning rate. Thus, individuals can be retarded or normal in developmental rate and retarded or normal in learning rate. Retardation in either realm will spell retardation as assessed by traditional intelligence tests, since these are a mixture of items that measure acquisition (e.g., vocabulary and general information subtests) and cognitive structures (e.g., problems involving logical reasoning). The 2 × 2 combinations indicated by this formulation suggest three possible classifications of familial retardates. Normal developmental rate and normal learning rate are both necessary for the manifestation of normal intelligence, as traditionally defined; neither alone is sufficient.

Our data pertain only to the relationship of MA to learning rate. No inferences are made here concerning the issue of developmental rate.

Jensen (1965) matched 40 institutionalized mentally retarded young adults (mean IQ = 58) with no known organic defects with 40 normal school children (mean IQ = 105) on MA (9 years). In both serial and paired-associate rote learning, the normal children had learning rates some 3 to 4 times faster, on the average, than the adult retardates. Furthermore, although there was no significant differ-

ence in the standard deviations for MA in the two groups, the retardates showed a significantly greater standard deviation of learning scores than the normals. The greater heterogeneity of learning rates of groups of retardates as compared with normals, when the groups are equally homogeneous in IQ and MA, was further substantiated in a study comparing learning rates in retarded, average, and gifted children (Jensen, 1963). There are evidently more ways of being retarded than of being either average or gifted in mental ability.

Rohwer (1967) compared a group of 48 institutionalized familially retarded adults with groups of normal children in Head Start and kindergarten and in Grades 1, 3, and 6 on paired-associate learning. The children were sampled from populations of low- and middle-socioeconomic status (SES). (The MA is close to the CA for the school children, but is slightly lower in the low-SES groups.) The results, shown in Figure 1, indicate that the average learning score of the retardates is significantly lower than that of any of the other groups as well as being significantly lower than all the other groups combined ($F = 103.22$, $df = 1/396$, $p < .01$). Comparison of the learning performance of the adult retardates and the middle-SES third graders is especially revealing, since the two groups have approximately the same MA (9.7 versus 9.6). Also, there was a larger standard deviation of learning scores in the retarded group than in any of the normal groups.

The relationship between learning rate and MA, at least in the mildly retarded (i.e., IQs of 50 to 75), is further complicated by socioeconomic status. Rapier (1968) closely matched Caucasian middle- and low-SES elementary school children ($N = 20$ in each group) in classes for the retarded on CA (124 months), MA (88 months), and IQ (70). None of the *S*s evinced any organic defects. The low-SES children showed consistently and significantly faster rates of paired-associate learning than the middle-SES children.

In view of the present results and consistent with our conceptualization, equivalence of developmental level need not

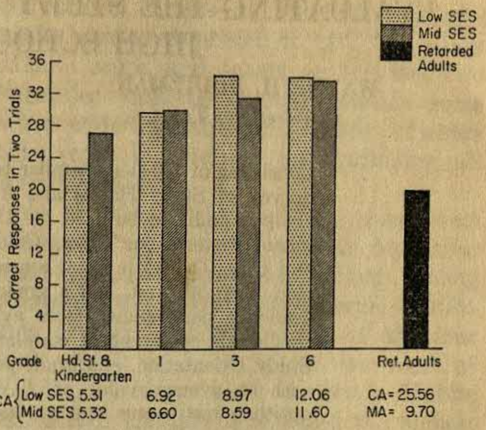


FIG. 1. Comparisons of low- and middle-socioeconomic groups of children at various grades in school with institutionalized retarded adults on paired-associate learning consisting of 24 picture pairs presented two times at a rate of 4 seconds per pair. $N = 48$ in each of the nine groups.

imply equality of performance on intellectual tasks, specifically, learning tasks. When equal-MA comparisons involve normals and familial retardates, differences in learning rate are to be expected, and, indeed, are found.

REFERENCES

JENSEN, A. R. Learning abilities in retarded, average, and gifted children. *Merrill-Palmer Quarterly*, 1963, **9**, 123–140.

JENSEN, A. R. Rote learning in retarded adults and normal children. *American Journal of Mental Deficiency*, 1965, **69**, 828–834.

RAPIER, J. The learning abilities of normal and retarded children as a function of social class. *Journal of Educational Psychology*, 1968, **59**, 102–110.

ROHWER, W. D., JR. On distinguishing the mentally retarded from the culturally disadvantaged. Paper presented at the meeting of the American Association of Mental Deficiency, Denver, May 1967.

WEIR, M. W. Mental retardation, technical comment. *Science*, 1967, **157**, 576.

WHITE, S. H. Evidence for a hierarchical arrangement of learning processes. In L. P. Lipsitt & C. C. Spiker (Eds.), *Advances in child development and behavior*. Vol. 2. New York: Academic Press, 1965.

ZIGLER, E. F. Familial mental retardation: A continuing dilemma. *Science*, 1967, **155**, 292. (a)

ZIGLER, E. F. Mental retardation, technical comment. *Science*, 1967, **157**, 578. (b)

(Received September 11, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 6, 404–409

EVALUATING THE STUDY HABITS AND ATTITUDES OF HIGH SCHOOL STUDENTS[1]

WAYNE H. HOLTZMAN AND WILLIAM F. BROWN
University of Texas *Southwest Texas State College*

Consisting of 100 items similar to those in the college version of the Survey of Study Habits and Attitudes (SSHA), Form H was developed and standardized on 11,218 students in Grades 7–12 from school systems in Texas, Colorado, Illinois, Maryland, Missouri, and Utah. Complete data, consisting of the SSHA, Form H, a scholastic aptitude test, and a grade-point average based on subsequent academic performance, were available for 10,888 students. Validity coefficients consisting of correlations between the total SSHA score, Study Orientation, and grade-point average ranged from .32 to .66, with an average value of .49. Correlations between the scholastic aptitude test scores and grades were only slightly higher, ranging from .19 to .83 with an average of .57. The low correlation between SSHA and scholastic aptitude (−.04–.54 with a mean of .27) indicated that the SSHA measures important traits related to school achievement that are untouched by the standard scholastic aptitude tests.

When first developed, the *Survey of Study Habits and Attitudes* (SSHA; Brown & Holtzman, 1953) was designated primarily for high school seniors and college freshmen. Consisting of 75 items, the original SSHA was standardized on samples comprised of nearly 4,500 freshmen enrolled in 11 different colleges and several hundred high school seniors. Scores on a suitable scholastic aptitude test and grade-point averages for the semester following administration of the SSHA were available to determine the validity of the SSHA for predicting academic success. The correlation between SSHA score and later grades ranged from .27 to .66 for men and from .26 to .65 for women. Consistently low correlations between the SSHA and scholastic aptitude made it possible to increase appreciably the prediction of grades by combining both scores (Brown & Holtzman, 1955; Holtzman, Brown, & Farquhar, 1954).

Successful application of the SSHA to a variety of problems ranging from identifying students who needed counseling to research on achievement motivation, created a demand for a similar instrument which could be given to children in junior and senior high schools. The Criticism of Education scale given to over 13,000 high school students in 1956 as part of the Texas Cooperative Youth Study (Moore & Holtzman, 1965) consisted of items like those in the SSHA dealing with scholastic motivation. A year later, an experimental version of the SSHA containing items suitable for junior high school was used successfully in a research program involving 1,470 seventh graders in four small cities (McGuire, Hindsman, King, & Jennings, 1963). Since revision of the SSHA was already underway to improve the questionnaire's usefulness for counseling purposes, the possibility of a parallel development to produce two forms, one for Grades 12–14 and the other for Grades 7–12, was particularly attractive. Form C of the SSHA for

[1] The research reported herein was supported in part by a grant from the Hogg Foundation for Mental Health, The University of Texas, Austin. The authors gratefully acknowledge substantial assistance from the following individuals who helped collect data from high school students across the country: William H. Dibrell, Francys Houston, and Glenn Campbell, San Marcos, Texas; Vernon Zunker, Sequin, Texas; James Williams, Glen Ellyn, Illinois; Dale Jackson, Austin, Texas; Catherine L. Beachley, Hagerstown, Maryland; Walter Bergmann, St. Louis, Missouri; Neida Clark, Durango, Colorado; Herbert Kaczmarek, Gunnison, Colorado; and Le Nora L. Losee, Salt Lake City, Utah. Data analysis was carried out with the assistance of Donald Witzke, Sara Currie, and Luis Laosa. Portions of this paper have been adapted from the *Survey of Study Habits and Attitudes Manual, Forms C and H*, written by the same authors and published by the Psychological Corporation, New York.

use at the college level has been described elsewhere (Brown & Holtzman, 1966). The present paper deals with the development of Form H for Grades 7–12 and research on its reliability and validity.

Development of the SSHA, Form H

Preliminary stages in the development of Form C for use with college students were completed before work on the high school version was begun. In addition to 70 items from the original SSHA, Form C contained 30 new statements which dealt largely with attitudes toward education and toward teachers. Enough progress had been made on the college revision to simplify development of Form H for use in junior and senior high schools. Operating independently, two committees of teachers undertook to revise items in Form C so they would conform with instructional procedures, academic requirements, and study conditions typical of Grades 7–12. A student committee then rephrased some of the items to eliminate wording likely to prove confusing to young teenagers. Of the 100 statements in Form C, only 17 had to be modified appreciably to maintain the same basic meaning.

The resulting version of the SSHA was then administered to all students in Grades 7–12 in San Marcos, Texas. Very few students had any difficulty in understanding and responding to the reworded statements. Similar results were obtained for Grades 7–9 in Livonia, Michigan by Morris (1961) who found a substantial correlation between SSHA scores and teachers' ratings of academic performance.

The 100 items in Form H were assigned to one of four subscales in exactly the same manner as they were classified in the revised college version, Form C (Brown & Holtzman, 1966), to preserve the parallel nature of the two forms. The original derivation of these four scales for Form C involved a priori classifications by 15 independent judges and item-subscale intercorrelations on a sample of 568 college freshmen. Containing 25 items each, the four subscales are Delay Avoidance, Work Methods, Teacher Approval, and Education Acceptance. Scores on the first two subscales can be combined to give the Study Habits score, and scores on the last two subscales yield the Study Attitude score when added together. The Study Orientation score is obtained by combining all four subscales.

As in the original SSHA, each item receives a weight of 0, 1, or 2, depending upon the empirically determined cutting points on the five-choice response continuum of each item. Since each of the four basic scales has a maximum raw score of 50, the maximum score on Study Habits and Study Attitude is 100 and the highest possible score for Study Orientation is 200. In the final published version of Form H, either IBM 805 or IBM 1230 answer sheets are available for machine scoring; special stencils are also provided for hand scoring of IBM 805 or 1230 answer sheets if desired (Brown & Holtzman, 1967). The Diagnostic Profile on the reverse side of the answer sheet facilitates interpretation and gives the counselor a convenient graphic record.

Reliability of Form H

Extensive studies of internal consistency and test-retest reliability for the four basic SSHA subscales as well as the Study Orientation score were first carried out for Form C, the college version of the SSHA. Using the Kuder-Richardson Formula 8 (Kuder & Richardson, 1937) for estimating test reliability from the variance of total scores and the sum of the item variances, coefficients ranging .87–.89 were obtained for the four basic subscales. A test-retest study over a 4-week interval with 144 freshmen who were given Form C yielded reliability coefficients of .93, .91, .88, and .90, respectively, for the Delay Avoidance, Work Methods, Teacher Approval, and Education Acceptance scales. The corresponding coefficients for a sample of 51 freshmen with a 14-week interval were .88, .86, .83, and .85, respectively. Since these results cannot be generalized with certainty to Form H, the high school edition, in spite of the high degree of similarity in the two forms, one additional study was conducted

using Form H with junior high school students.

A sample of 237 ninth graders in San Marcos High School was given the SSHA twice, with an interval of 4 weeks between sessions. The test-retest reliability coefficients were .95, .93, .93, and .94, respectively, for the Delay Avoidance, Work Methods, Teacher Approval, and Education Acceptance scales. The means and standard deviations remained essentially unchanged over the 4-week period.

The stability of Form H scores for ninth graders compares favorably with the stability of Form C scores for college freshmen. These studies indicate that the four subscales are sufficiently stable through time to justify their use in predicting future behavior or in assessing the degree of change in study habits and attitudes after counseling.

Standardization and Validation

Preliminary standardization of the SSHA, Form H, was carried out in the fall semester, 1964, on 3,731 students in Grades 7–12 in 10 junior and senior high schools located in small towns throughout central Texas. Correlations between the Study Orientation score and grade-point averages at the end of the year ranged from .31–.85, sufficiently high to justify a standardization program on a national scale.

Arrangements were made to collect data during the fall semester, 1965, from students in Grades 7–12 in Austin, Texas; Durango, Colorado; Glen Ellyn, Illinois; Gunnison, Colorado; Hagerstown, Maryland; Salt Lake City, Utah; and St. Louis, Missouri. Scores on an acceptable scholastic aptitude test and subsequent grade-point averages for the fall semester were obtained for nearly all of the students tested. When added to the central Texas samples studied the year before, the new SSHA protocols yielded a total of 11,218 students upon which percentile norms have been developed—5,425 cases for junior high school norms (Grades 7–9) and 5,793 cases for senior high school norms. Complete data, consisting of the SSHA, a scholastic aptitude test, and a grade-point average, were available for 10,888 students.

Since each school system had its own policies concerning the kind of aptitude test to be generally administered, either percentile scores or IQ equivalents were used in the statistical analysis. The scholastic aptitude tests included the Cooperative School and College Ability Tests, the Differential Aptitude Tests (Form L), the Henmon-Nelson Tests of Mental Ability, the Iowa Tests of Educational Development, the Lorge-Thorndike Intelligence Tests, the Otis Quick-Scoring Mental Ability Tests (Beta Test, Form CM, and Gamma Test, Form AM), the Pintner General Ability Tests, and the Preliminary Scholastic Aptitude Test.

The criterion of school achievement, grade-point average for the fall semester, was generally obtained by assigning weights of 4, 3, 2, 1, and 0 to grades of A, B, C, D, and F, respectively. Only courses in the so-called "solids," that is, mathematics, science, social studies, foreign language, and English, were used in computing grade-point averages.

The intercorrelations of the seven SSHA scores, the scholastic aptitude test score, and the grade-point average were computed separately for each grade and school system. The detailed results of this analysis are given in the *Survey of Study Habits and Attitudes Manual, Forms C and H* (Brown & Holtzman, 1967). Only a summary of the findings can be presented here.

All 49 of the correlations between the total SSHA score, Study Orientation, and grade-point average proved to be highly significant. The individual validity coefficients ranged from .32 to .66; the average correlation based on all 10,888 cases was .49.

Correlations between the scholastic aptitude test score and grade-point average were only slightly higher, ranging .19–.83 with an average of .57. The correlation between the SSHA and the scholastic aptitude test was generally low, ranging −.04–.54 with a mean of .27, suggesting that the SSHA measures important traits related

TABLE 1

Mean Correlation of Survey of Study Habits and Attitudes (SSHA), Form H; Scholastic Aptitutude (SA); and Grade-Point Average (GPA) Together with Multiple (R) and Partial (r) Correlations of Scores with Grade-Point Average

Grade	N	SSHA (1) with GPA (3)	SSHA (1) with SA test (2)	SA test (2) with GPA (3)	$R_{3.12}$	$r_{31.2}$	SSHA total M	Study Orientation score SD
7	1,684	.55	.32	.61	.72	.47	106.5	33.1
8	1,628	.52	.29	.59	.69	.45	107.0	32.8
9	2,005	.49	.29	.52	.63	.41	101.6	31.5
10	2,064	.49	.29	.62	.70	.41	97.9	31.6
11	1,840	.47	.24	.57	.67	.42	97.7	30.9
12	1,667	.46	.20	.53	.66	.43	102.9	30.3
Total	10,888	.49	.27	.57	.67	.43		

Note.—Mean correlations were obtained by converting each r into its Fisher's z function, weighting by the appropriate number of cases, averaging the values, then reconverting. Multiple and partial coefficients were derived from the weighted averages.

to school achievement that are untouched by the standard scholastic aptitude tests. Very similar findings were obtained among college students with both Form C and the original version of the SSHA.

Further insight into the nature of the relationship between the study habits and attitudes, scholastic aptitude, and school achievement can be gained by inspection of the multiple and partial correlations using grade-point average as a criterion. These statistics, together with the mean zero-order correlations among the three variables, are presented separately for each grade in Table 1. It is apparent in every case that both the SSHA and the scholastic aptitude tests contribute in significant and distinctly different ways to the successful prediction of actual school achievement. The multiple correlation using both predictors is appreciably higher than the correlation of either one alone with the criterion. Because of the low intercorrelation between the SSHA and scholastic aptitude, the correlation between the SSHA and grade-point average with scholastic aptitude partialed out is still quite high, ranging .41–.47 across the six grade levels.

Table 1 also contains the mean and standard deviation of the Study Orientation score for each of the school grades. The slight but regular drop in variance with increasing grade level, coupled with a similar drop in the correlation between SSHA and grade-point average, is probably due to gradual loss of extremely poor students who drop out of high school before finishing. Reasons for the minor fluctuation in mean scores, however, are unknown.

Unlike the original version of the SSHA, both Form C and Form H consist of four basic subscales, each containing 25 items clustered together in a scale because of commonly shared content. It is of some interest to examine the intercorrelations of these four scales as well as their relative value in predicting the criterion of school

TABLE 2

Average Intercorrelation Coefficients, Means, and Standard Deviations of Basic Scores on the Survey of Study Habits and Attitudes, Form H

Scale	DA	WM	TA	EA
Delay Avoidance (DA)		.70	.51	.65
Work Methods (WM)			.56	.65
Teacher Approval (TA)				.75
Education Acceptance (EA)				
M	21.9	22.3	28.5	27.5
SD	9.7	9.2	10.1	8.7

Note.—N = 11,218.

achievement. Table 2 contains the mean intercorrelations of the four basic scales. These correlation coefficients were obtained by converting each entry in a table of correlations for each of the grades at each participating school system to its Fisher's z function, weighting each by its appropriate number of cases, averaging the values, and reconverting. The values as presented are based on all 11,218 cases.

Intercorrelations among the four subscales are moderately high and positive, suggesting that one major dimension running through all four scales is sufficient to account for most of the variance. The Study Orientation score obtained by summing the raw scores on the four basic scales is the best single measure of this dimension. It should be noted, however, that the highest correlations, .70 and .75, occur between the pairs of scales making up the derived scales of Study Habits and Study Attitudes, respectively. Given the large sample size and resulting high stability of the obtained correlations, it can be concluded that Delay Avoidance and Work Methods have more in common with each other than either has with the remaining two scales. The same can be said for Teacher Approval and Education Acceptance. This finding provides some empirical justification for the derived scales based on these two pairs of subscales.

Correlations between each of the four basic subscales, grade-point average, and the scholastic aptitude measures are given in Table 3. Delay Avoidance shows the least amount of overlap with scholastic aptitude (r = .16) while Work Methods shows the most (r = .37). Validity coefficients for the four scales are only slightly lower than the comparable coefficient for the total SSHA score, Study Orientation. The values of .47 and .48 for Work Methods and Education Acceptance are almost identical with the correlation of .49 for Study Orientation.

TABLE 3

SURVEY OF STUDY HABITS AND ATTITUDES (FORM H) SUBSCALE CORRELATIONS WITH GRADE-POINT AVERAGE (GPA) AND SCHOLASTIC APTITUDE (SA)

Scale	r with GPA[a]	r with SA test[b]
Delay Avoidance	.41	.16
Work Methods	.47	.37
Teacher Approval	.35	.29
Education Acceptance	.48	.26

Note.—Subscale correlations were obtained by converting each r into its Fisher's z function, weighting by the appropriate number of cases, averaging the values, then reconverting.

[a] N = 10,888.

[b] N = 7,157.

DISCUSSION

The above results clearly demonstrate the reliability and validity of the SSHA, Form H, when extended downward as far as the seventh grade. Correlations between the high school version of the SSHA and subsequent academic grades are even higher than those generally obtained for the college edition, Form C. The average correlation between SSHA, Form C, and grades for freshmen in six different colleges was .36 (Brown & Holtzman, 1967), as compared to .49 for SSHA, Form H, and high school students. Undoubtedly, much of this difference is due to the greater heterogeneity of the high school students. The mean score on Study Orientation for Form H is 100.3, as contrasted to a mean of 114.2 for Form C, indicating that the greater heterogeneity is due largely to a higher proportion of low scores among high school students than among college freshmen.

Granted that the SSHA is a useful instrument for evaluating the study habits and motivation of students in Grades 7–12 as well as in college, to what extent can the habits and attitudes of such students be improved by special training or counseling? A recent experiment by Haslam and Brown (1968) indicates that substantial improvement can be obtained from appropriate study-skills instruction—improvement not only in SSHA scores but also in subsequent academic grades, as compared to matched control cases. Clearly, the junior or senior high school offers an even more meaningful opportunity for systematic efforts aimed at improving study

habits, attitudes, and motivation than does the freshman year in college.

In spite of the substantial overlap in meaning among scores from the four basic SSHA subscales, the use of subscales has real value in individual counseling. Because it is hard to get across the content of the survey in a simple dramatic manner using the original SSHA single score, the use of subscale scores rather than individual items makes it possible for a counselor to stress four somewhat different areas rather than a few individual items. In this way the student can profit most from such counseling by being able to remember where his difficulties lie. In addition, individuals who wish to engage in evaluative research may more profitably use the four relatively homogeneous subscale scores rather than either single items or only one general score.

REFERENCES

Brown, W. F., & Holtzman, W. H. *Survey of Study Habits and Attitudes*. New York: Psychological Corporation, 1953.

Brown, W. F., & Holtzman, W. H. A study-attitudes questionnaire for predicting academic success. *Journal of Education Psychology*, 1955, **46**, 75–84.

Brown, W. F., & Holtzman, W. H. *Survey of Study Habits and Attitudes manual, Form C.* New York: Psychological Corporation, 1966.

Brown, W. F., & Holtzman, W. H. *Survey of Study Habits and Attitudes manual, Forms C and H.* New York: Psychological Corporation, 1967.

Haslam, W. L., & Brown, W. F. Effectiveness of study-skills instruction for high school sophomores. *Journal of Educational Psychology*, 1968, **59**, 223–226.

Holtzman, W. H., Brown, W. F., & Farquhar, W. G. The Survey of Study Habits and Attitudes: A new instrument for prediction of academic success. *Educational and Psychological Measurement*, 1954, **14**, 726–732.

Kuder, G. F., & Richardson, M. W. The theory of the estimation of test reliability. *Psychometricka*, 1937, **2**, 151–160.

McGuire, C., Hindsman, E., King, F. J., & Jennings, E. Dimensions of talented behavior. *Educational and Psychological Measurement*, 1963, **21**, 3–38.

Moore, B. M., & Holtzman, W. H. *Tomorrow's parents*. Austin: University of Texas Press, 1965.

Morris, F. L. The validity of the Brown-Holtzman Survey of Study Habits and Attitudes 1960 experimental revision for grades 7–12. Unpublished master's thesis, Wayne State University, 1961.

(Received October 24, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 6, 410–413

EFFECTS OF INCIDENTAL CUES AND ENCODING STRATEGIES ON PAIRED-ASSOCIATE LEARNING[1]

JOSEPH R. JENKINS
University of Delaware

Several studies have demonstrated that picture stimuli relative to word stimuli facilitate paired-associate learning. To determine whether incidental cues in pictures, or possibly a different encoding strategy elicited by pictures, produce this learning difference, 6 treatments, involving 72 *Ss*, were devised. When incidental cues in pictures were varied by presenting *Ss* with several pictures representing the same concept, pictures unexpectedly produced faster learning than words ($p < .001$). Adding incidental cues to word stimuli did not increase learning efficiency relative to normally presented words. Moreover, requiring *Ss* to label the stimuli, pictures or words, failed to produce a significant difference in performance from that of *Ss* who learned under standard conditions. Finally, superiority of picture stimuli over word stimuli was replicated ($p < .01$).

Several investigators (Deno, 1968; Lumsdaine, 1949; Pavio & Yarmey, 1966) have demonstrated that picture stimuli, compared to word stimuli, facilitate paired-associate (PA) learning. Others (Jenkins, Neale, & Deno, 1967) who measured recognition memory for both media found that pictures are more easily recognized than words. The present study attempts to determine the characteristics of pictorial representation that produce the observed differences in learning and memory.

Two alternative explanations are proposed to account for the effectiveness of pictorial stimuli in learning. An incidental cue explanation postulates a difference in the degree of perceptual richness for words and pictures. In the typical PA experiment, words and pictures are chosen so that the picture and word unequivocally represent the same concept. Generally, this is accomplished by choosing pictures that are given a common verbal label by nearly all *Ss*. This procedure results in a set of pictures that are presumably equivalent to a corresponding set of words in the extent to which they elicit some common verbal response. An obvious fact, however, is that the stimuli, pictures and words, contain cues that are additional or incidental in the sense that these cues may be altered without changing the tendency of *Ss* to name the stimulus appropriately. A cue is therefore incidental if changing it does not reduce the probability that *Ss* will give the intended label. For example, incidental cues in the picture stimulus, BOY, may be distinctive clothing, a facial expression, or books the boy is carrying. On the other hand, incidental cues, although present in the word BOY, may be less conspicuous, such as pica type, all upper-case letters, and context. Moreover, the fact that other word stimuli may share a number of these same attributes increases the similarity among the stimuli.

It may be, then, that incidental cues produce the picture-word difference in learning by presenting additional "pegs" (functional stimuli) to which learners can attach the to-be-learned response, by making the stimuli physically more dissimilar and thus more discriminable, or by changing the stimulus on an abstraction dimension.

Another explanation investigated in the study involves a difference in encoding processes for the two types of stimuli.

[1] This research was supported by grants to the University of Minnesota, Center for Research in Human Learning, from the National Science Foundation (OS541), the National Institute of Child Health and Human Development (5-P01-HD-01136-03), and the Graduate School of the University of Minnesota. The data on which this paper is based were included in the author's dissertation presented in partial fulfillment of the requirements for the Doctor of Philosophy degree at the University of Minnesota. The author expresses gratitude to Daniel C. Neale, who directed this study.

Learners may have a weaker tendency to represent or encode a stimulus verbally when it is a picture rather than a word (Lumsdaine, 1949). Bousfield (1961) contended that associative interference occurs in learning as a result of similarity among representational responses. Deno (1968) recently reported that the largest picture-word differences in learning occur with conceptually similar stimuli, that is, stimuli which presumably elicit associatively similar representational responses. Deno's study suggests that by representing the stimulus picture nonverbally, the learner reduces associative interference that he would otherwise experience had he represented the stimulus verbally.

The present study seeks to assess the contribution of incidental cues to efficiency in PA learning by maximizing or minimizing their presence and noting the effect on learning. This was attempted by presenting *S*s with several different pictures representing the same concept (washing out incidental cues) or by presenting word stimuli to which incidental cues, such as, color, size, and unique lettering, were added (enriching words). In addition, the study was designed to bear upon the proposition that picture-word learning differences are produced by different encoding strategies. It was hypothesized that increasing the similarity of the strategies used to encode words and pictures increases similarity in performance. This was tested by noting if the requirements to label all stimuli overtly produced a decrement in performance, particularly when the stimuli were pictures.

METHOD

Subjects

Subjects for this experiment were 72 students enrolled in introductory psychology classes at the University of Minnesota. Each student's participation was voluntary but was also rewarded with course credit.

Several restrictions were placed on the selection of *S*s. In an effort to decrease the variability among *S*s, only female *S*s were allowed to participate. Secondly, to insure that all *S*s possessed approximately the same language habits, only those who spoke English natively were allowed to participate. Finally, since Japanese words were employed as responses, no *S* could participate who had previously studied an oriental language. Twelve *S*s were assigned randomly to each of six experimental groups.

Materials

The *S*s learned a PA list consisting of 12 stimulus-response pairs. The stimuli were either words, words with incidental cues added, or pictures representing 12 common objects. Responses were 12 Japanese words.

The 12 stimuli were conceptually similar, that is, there were three instances from each of four conceptual categories—animal, clothing, furniture, and people. Each category instance was maximally separated from other instances in the same category. Such a list was chosen on the basis of Deno's (1968) finding that conceptually similar, maximally separated stimuli produce the greatest picture-word difference in learning, since it is this difference that the current study seeks to investigate. The following stimulus concepts and responses were used.

Stimulus	*Response*
TABLE	ATSUI
MAN	HUNE
HAT	RIKO
DOG	KARAI
CHAIR	BAKA
BOY	AMAI
COAT	HIKUI
CAT	HAYAI
BED	KURO
GIRL	TAKO
TIE	CHIKAI
MOUSE	TOOI

The same 12 Japanese words were used as responses regardless of the experimental condition to allow for controlled comparisons among conditions. In addition, two random compositions of stimuli and responses were used so that one-half of the *S*s in each experimental condition received a different random order. In the two random pairings stimuli were never paired with the same response.

Three random orders of stimuli were obtained to guard against sequential response learning. This was accomplished, first, by ordering the concepts within a list and, second, by assigning the instances within each category. For a given *S* the same stimulus concept was always paired with the same response term.

Apparatus

The picture stimuli and responses were photographed on 35mm. black-and-white film, and negatives were slide-mounted. A typical 2 second–2 second PA anticipation method was employed.

Procedure

All *S*s participated individually. Each *S*, upon arrival at the laboratory, was randomly assigned to one of the following six experimental conditions.

Word-Normal (*WN*). The *S*s learned verbal responses to word stimuli in normal orthography (typewriter capitals). A sample list of this condition is given above.

Picture-Normal (*PN*). The *S*s learned verbal responses to pictorial stimuli. The same picture was always used as a particular stimulus.

Word-Enriched (*WE*). The *S*s were shown "perceptually enriched" words, that is, the stimulus words were reproduced with distinctive lettering, colors, and slants. Each stimulus was enhanced by the addition of incidental cues.

Picture-Washout (*PWa*). The *S*s learned responses to picture stimuli, but the same picture never occurred on two successive trials. For example, if on the first trial a picture of a boy was paired with the Japanese word KARAI, on the second, third, fourth, and fifth trials a different picture of a boy was paired with the Japanese word KARAI. Each of the 12 stimuli was represented by five different pictures. All the pictures were highly labelable.

Word-Label (*WL*). The *S*s were required to label the stimulus word overtly when it appeared. Word stimuli were in normal form.

Picture-Label (*PL*). The *S*s were instructed to label each picture overtly as it appeared. Picture stimuli were in normal form.

After *E* described the learning task to each *S*, he proceeded to give one familiarization trial with the stimuli and three familiarization trials with the responses. The familiarization procedure was intended to reduce variability in performance that is ordinarily attributed to response learning. Since different groups required longer to learn, the sum of correct responses over the first seven trials, rather than the total number of correct responses, was used to measure group performance.

RESULTS

The performance and tests of significance are presented in Table 1.

The dependent variable, the mean number of correct responses, was employed in five prior comparisons that follow. To test the incidental cue hypothesis, performance of Group WN was tested, first, against PWa and, second, against WE.

The results of the first comparison fail to support the incidental cue hypothesis. Not only did washing out incidental cues fail to produce comparable performance between Groups PWa and WN (t = 4.53, $p < .001$), but also, conversely, Group PWa seemed to have learned as easily as group PN.

TABLE 1

MEANS AND STANDARD DEVIATIONS ON THE NUMBERS OF CORRECT RESPONSES AND PERFORMANCE COMPARISONS FOR THE TREATMENT GROUPS

Group	*M*	*SD*
Picture stimuli		
Picture-Washout (PWa)	51.0	9.75
Picture-Normal (PN)	43.83	7.69
Picture-Label (PL)	38.42	15.58
Word stimuli		
Word-Enriched (WE)	34.25	15.13
Word-Label (WL)	34.17	10.65
Word-Normal (WN)	25.92	16.54

Comparison	*t* ratio
PWa versus WN	4.53**
WE versus WN	1.29 *ns*
PN + PL versus WN + WL	2.92*
PL + WL versus PN + WN	0.38 *ns*
PN − WN versus PL − WL	1.80 *ns*

* $p < .01$.
** $p < .001$.

The second comparison, similarily, fails to support the incidental cue hypothesis, in that enriching words did not significantly facilitate learning relative to normally written words ($t = 1.29$, *ns*).

The third comparison of the rate of learning[2] with picture stimuli as opposed to learning with word stimuli supports the common finding that picture stimuli facilitate learning relative to word stimuli (t = 2.92, $p < .01$).

The fourth comparison tests the effect on learning of requiring *S*s to label both picture and word stimuli. Results suggest that *S*s who label the stimuli overtly do not perform with significantly different success (t = .38, *ns*) than *S*s who are not required to label the stimuli.

Finally, the interaction of stimulus mode with a requirement to label, the fifth comparison, is not statistically significant ($t = 1.8$, *ns*).

DISCUSSION

The four conditions involved in the test of the incidental cue hypothesis are PN,

[2] Note that the third, fourth, and fifth comparisons constitute the treatment and interaction effects of a normal 2 × 2 analysis of variance.

WN, PWa, and WE. The hypothesis states that differences in perceptual richness between words and pictures produce the observed differences in learning. Therefore, raising the salience of incidental cues by embellishing the words should facilitate PA learning relative to normally written words. Also, washing out incidental cues in pictures by varying the pictures on acquisition trials should produce performance comparable to that produced by normally presented words. Contrary to predictions made on the basis of the incidental cue hypothesis, embellishment of word stimuli did not facilitate learning relative to normally produced words; and minimizing incidental cues in picture stimuli failed to reduce learning efficiency to a level comparable to normally presented words. Furthermore, an explanation for picture superiority in PA learning apparently does not involve a difference in the linguistic encoding of the stimuli, since requiring learners to label the stimuli did not significantly affect performance.

A psychological explanation for the differential effectiveness of word and picture stimuli in PA learning is not immediately apparent. Perhaps an alternative explanation along the lines of Kagan's (1967) attentional determinants of learning is worthy of investigation. Kagan presented evidence to suggest that the attention given to and the reinforcement value of creating schema (the psychological representation of an external pattern) is inversely related to the predictability of the external pattern. With this consideration, pictorial representations typically employed in comparisons of stimulus mode may very likely be expected to elicit greater attention from the learner, because they fail to correspond with his own schematic representation of the concept. This would certainly be the case under conditions where a single concept is represented by a series of different pictures (as in PWa). On the other hand, a single word, regardless of its presentation, should not influence the natural schematic representation of the concept.

Clearly, simplistic notions involving incidental cues and labeling strategies must be reexamined. The results of the present investigation suggest that a learner presented with a picture stimulus adopts as his functional stimulus some aspect of the stimulus configuration, perhaps a schema, the efficacy of which is independent of the verbal label for that stimulus and of incidental cues present in the stimulus.

REFERENCES

BOUSFIELD, W. A. The problem of meaning in verbal learning. In C. N. Cofer (Ed.), *Verbal learning and verbal behavior.* New York: McGraw-Hill, 1961.

DENO, S. L. Effects of words and pictures as stimuli in learning language equivalents. *Journal of Educational Psychology,* 1968, **59,** 202–206.

JENKINS, J. R., NEALE, D. C., & DENO, S. L. Recognition memory for word and picture stimuli. *Journal of Educational Psychology,* 1967, **58,** 303–307.

LUMSDAINE, A. A. The effectiveness of pictures versus printed words in learning simple verbal associations. Unpublished doctoral dissertation, Stanford University, 1949.

KAGAN, J. On the need for relativism. *American Psychologist,* 1967, **22,** 131–142.

PAVIO, A., & YARMEY, A. D. Pictures vs. words as stimuli and responses in paired-associate learning. *Psychonomic Science,* 1966, **5,** 235–236.

(Received November 15, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 6, 414-419

CLASSROOM CLIMATE AND INDIVIDUAL LEARNING[1]

HERBERT J. WALBERG AND GARY J. ANDERSON
Harvard University

The research is one of a series testing the Getzels-Thelen theory of the classroom as a social system. From a new measure of student perception of classroom climate, 18 subscores were obtained in November from 76 classrooms throughout the United States and used to predict 9 congitive, affective, and behavioral measures of learning at the end of the school year (regression-adjusted for initial differences). More than 4 times as many correlations as the chance expectancy were significant ($p < .05$). Among the structural and affective climate measures, variables grouped under the rubrics "isomorphism," "organization," and "synergism" predicted learning variables more frequently than "coaction" and "syntality."

Two recent studies have shown that scores obtained on a measure of the socioemotional climate of the classroom (Walberg, 1966) can be predicted from earlier measures of (*a*) teacher personality (Walberg, 1968a) and (*b*) student ability and interest in the subject (Walberg & Anderson, 1968). Yet this work is incomplete in that it does not demonstrate that the student's individual satisfaction with the climate of the class makes for learning, the criterion of institutional effectiveness espoused by school boards, parents, administrators, and teachers. The intent of the present research is to investigate this crucial relationship and to explore empirically further hypotheses derived from a sociopsychological theory of the classroom as a social system (Getzels & Thelen, 1960).

Getzels and Thelen make an analytic distinction between institutional role expectations and individual personality dispositions, which both bear upon the climate of the class. The consellation of role expectations can be termed the "structural" dimension (Walberg, 1968b); it refers to the structure or organization of student roles within the class, for example, such things as goal direction and democratic policy. The structural dimension applies to shared, group-sanctioned classroom behavior, while the "affective" dimension pertains to idiosyncratic personal dispositions to act in a given way to satisfy individual personality needs. Aspects of the affective dimension are such things as satisfaction, intimacy, and friction in the class. A recent multivariate study, in the same theoretical vein, of 72 classrooms showed that student perceptions of the structural and affective aspects of socioemotional climate are strongly related (canonical correlations as high as .8). And although the patterns of correlation are complex, they are interpretable in terms of the Getzels-Thelen conceptual scheme and certain other sociopsychological theories (Walberg, 1968b). The present study fits into the series as follows:

Input September pretests	Throughput December midtest	Output May posttests
Personality and cognitive measures on	Classroom climate	Student learning
Teachers ⟶	Structure ⟶ ↕	Cognitive Affective
Students ⟶	Affective ⟶	Behavioral

The solid lines refer to relationships that have already been established in prior work; the broken lines refer to the purpose of this study: the examination of the hypothesis that individual student achievement and interest in the subject at the end of the school year can be predicted from

[1] This research is part of the evaluation of Harvard Project Physics, a course-development project supported by the Carnegie Corporation of New York, the National Science Foundation, the Sloan Foundation, and the United States Office of Education. The authors thank Fletcher G. Watson and Wayne W. Welch for comments on a draft of the manuscript and Mary Hyde and Arthur Rothman for computer consultation and special programing. The second author is now at McGill University.

structural and affective aspects of classroom climate measured at midyear.

Previous studies were of the class as a whole and, hence, the units of analysis were the means of the student measures within each class. It has been shown that the correlation of means of subgroups and the correlation of individuals within the same sample can differ in sign and magnitude (Robinson, 1950). Hence, a parallel means analysis is in progress to complete the series above, but the focus of this study is the individual. It seeks to determine the learning of individuals with different perceptions of classroom climate rather than the mean perception of entire classes.

METHOD

Subjects and Instruments

Some 2100 high school juniors and seniors in 76 classes throughout the country participated in the preliminary evaluation of Harvard Project Physics, an experimental course using a variety of new instructional media and emphasizing the philosophical, historical, and humanistic aspects of physics. The mean Henmon-Nelson IQ of a random sample of the group is 115. Their scores on five instruments constitute the data for analysis.

The battery of cognitive, affective, and behavioral criterion measures includes the Physics Achievement Test, the Science Process Inventory, the Semantic Differential for Science Students, and the Pupil Activity Inventory. The Physics Achievement Test (Ahlgren, Walberg, & Welch, unpublished, 1966) is a 36-item multiple-choice test designed to measure general knowledge of physics. It has a Kuder-Richardson formula 20 reliability (Guilford, 1954) of .76 based on a random sample of 400 high school students at the end of their physics course. The Science Process Inventory (Welch & Pella, 1967) consists of 100 true-false statements describing the assumptions, activities, products, and ethics of science. The test was validated on a sample of eminent scientists and has a Kuder-Richardson formula 20 reliability of .86.

The Semantic Differential is familiar to many researchers and has been described elsewhere (Osgood, Suci, & Tannenbaum, 1957; Walberg & Anderson, 1968). Six clusters reflecting affective objectives of Harvard Project Physics were selected for analysis. Using the Spearman-Brown formula to correct the mean item-intercorrelations for the number of items (Guilford, 1954) yields reliabilities of about .8. (See Table 1 for reliabilities of all scales and tests discussed here.)

The Pupil Activity Inventory was described by Cooley and Reed (1961). It consists of a number of adolescent science activities, and the student is asked to indicate the frequency of his participation in each. Walberg (1967) re-factor analyzed the instrument for the present sample and found five dimensions: Academic, Biological, Tinkering, Cosmology, and Applied Life. The Academic, Tinkering, and Cosmology cluster scores were summed for a physics activity score which yields a Spearman-Brown corrected internal consistency of .76.

The first form of the Classroom Climate Questionnaire (Walberg, 1966) consists of 80 items describing characteristics of school classes, for example, "The class members are working toward many different goals." The respondent expresses agreement or disagreement with each on a 4-point scale. The instrument yields 18 factor-analytically derived cluster scores which, for individuals, range in corrected split-half reliability from .41 to .86 (See Table 1 and Walberg & Anderson, 1968). A revised instrument with more items and, hopefully, greater reliability is being used to replicate the work this year.

Procedure

The data were obtained using a randomized data-collection system within each class which tends to minimize individual testing time and maximize the number of tests which can be administered (Walberg & Welch, 1967). The system is most appropriate for class means analysis, but it does provide patterns of scores for studies of individual students as well, with certain restrictions. Random halves of the group of students took the criterion measures at the beginning and at the end of the year; a random fourth took the Classroom Climate Questionnaire at midyear. The sampling fraction for any combination of tests is the product of the sampling fractions for the combination. Thus, for the midtest and any posttest, a fourth times a half or an eighth took both measures. To bring pretests into the analysis, the eighth must be multiplied by a half, giving one-sixteenth. Thus, for a total of 1700 students who finished the course, about 214 took the midtest and a given posttest; and 106 took the same pretest and posttest, as well as midtest. Actually, because of absentees and unusable answer sheets, the figure is about 85 for any given combination of pre-, mid-, and posttest.

From the group of 25 subscores on the tests given at the beginning and at the end of the course, 9 were selected as criteria for measuring student learning, since they measure cognitive, affective, and behavioral course objectives. They are: physics achievement; science understanding; six semantic differential measures; and physics activities, which is the sum of the Academic Science, Cosmology, and Tinkering scales on the Pupil Activity Inventory. The reliabilities of the scales are shown in Table 1. Using a method described by Ferguson (1959), regression-adjusted gains ("delta") scores (the posttests' standardized deviations from predicted scores based on the

TABLE 1

CORRELATIONS OF CLASSROOM-CLIMATE AND STUDENT-LEARNING MEASURES

Classroom climate	Cognitive		Affective[c]						Behavioral
			Laboratory		Universe		Physics		
	Physics achievement[a] (76)	Science understanding[b] (86)	Important (85)	Fun (71)	Beautiful (65)	Friendly (61)	Interesting (81)	Important (69)	Physics activities[d] (70)
Structural aspects									
Coaction									
Subservient (57)									
Strict control (51)									−21
Speech constraint (41)						−25*			
Isomorphism									
Democratic (80)			26*	28*	22*	18			
Stratified (55)		−22*	−34*	−22*	−25*	−24*	−27*		
Egalitarian (67)		20			32*	40*			
Organization									
Goal direction (80)			41*	25*		26*		24*	
Disorganized (55)		−21					−30*		
Formality (51)			23*		23*				
Goal diversity (64)	−19*								
Affective aspects									
Syntality									
Classroom intimacy (79)									
Alienation (75)									−20
Group status (68)									
Synergism									
Satisfaction (53)			30*	24*	18			20	
Friction (86)		−31*	−24*		−23*	−27*		−23*	
Personal intimacy (58)	20*		20						26*
Miscellaneous									
Social heterogeneity (79)	−23*				−27*				
Interest heterogeneity (51)		21	−23*					−19	

Note.—Decimals and correlations below the .10 significance are omitted. Test reliabilities are given in parentheses, decimals omitted.

[a] $n = 96$.
[b] $n = 76$.
[c] $n = 82$.
[d] $n = 82$.
* $p < .05$.

pretest), were calculated for each of the criteria. These scores represent the student's learning on each criterion during the course adjusted for initial status. The adjusted criteria were correlated with each of the 18 measures of classroom climate.

RESULTS

Table 1 contains 32 statistically significant correlations ($p < .05$) between measured perceptions of classroom climate and the adjusted learning variables. This amounts to four times the chance expectancy in a 9×18 matrix of 162 elements.[2] The estimates of association are conservative since the criterion-test scores are not highly reliable, and adjusted gain scores are even less reliable. Using the conservative attenuation correction for un-

[2] Stepwise multiple correlations were also calculated with significant results accounting for up to 40% of the uncorrected variance in the learning criteria with three predictors. However, because of the small number of cases, the uncertainty of the stepwise procedure without cross-validation, and great number of beta weights, the results are not reported here.

reliability of the criterion test only, the correlations rise from 7% to 28%. The correlations rise from 16% to 200% when corrected for criterion and predictor unreliability. Since the question of reliability of gain scores is still unsettled (see Harris, 1963), this further correction for a third source of error variance is not considered here. In any case, uncorrected correlations and scale reliabilities are shown in Table 1. The interested reader is referred to Guilford (1954) for attenuation-correction formulas.

Discussion

While the study is exploratory and employs a preliminary form of the instrument, the results are statistically significant and meaningful enough to warrant interpretation. One way to do this is to characterize the perception of classroom climate for students who made greatest gains on the different criteria by examining the columns of correlations in Table 1. Students who gained the most on the Physics Achievement Test, for example, perceived their classes as socially homogeneous, intimate groups working on one goal; one might speculate that the goal is high achievement on physics tests. On the other hand, students who grew more in science understanding saw their classes as well organized with little friction between their fellow students, and although the class is seen as egalitarian and unstratified, the students had a greater variety of interests. Thus, different perceptions of classroom climates are associated with different kinds of cognitive growth—achievement and science understanding.

Perceptions of climate also predict the affective growth the course is intended to bring about. The correlates of only one of the two ratings for each of the three concepts reported in Table 1 are discussed here. Students who reported greater enjoyment of laboratory work perceived their classes as unstratified, democratic in policy setting, having a clear idea of class goals, and satisfying. Students who gained the most interest in physics saw their classes as well organized and unstratified. Those who rated the concept Universe more friendly saw their classes as having clear goals, democratic in policy setting, egalitarian, unstratified, and as having less internal friction and speech constraint. Finally, students who reported engaging in more physics activities, because they were interested, felt more personally intimate with their fellow class members, less alienated, and less strictly controlled.

Thus, students with various perceptions of classroom climate grow in different ways during a course. Another way of examining the results is to analyze the correlations across the rows to determine which climate variables correlate most often with student learning variables. The structural climate variables can be divided into three subgroups: those having to do with "coaction," "isomorphism," and "organization."[3] An enormous amount of research has investigated "teacher-centered" versus "student-centered" classrooms or other variations on the themes of "authoritarian" and "dominant" teaching methods (Gage, 1963). However, most of the studies, whether they employ tabulations of systematic observations or observer ratings, fail to significantly account for variance in student learning. Three "coaction" climate variables—subservient, strict control, and speech constraint—seem to be related to this dimension; and among the three, there is only one correlation with student learning.

On the other hand, a more promising dimension for predicting learning is "isomorphism," or the perceived equality of class members. Democratic, stratified, and egalitarian correlate significantly with learning in 11 instances. Stratification correlates with six learning measures, more

[3] These terms are used as a matter of convenience in discussing the results. Except in the case of "syntality," a term employed for some years by Cattell (see Bereiter, 1966), the authors refrain from using or adding new terms to the copious jargon of psychology. In all other cases, the authors have used words with dictionary definitions which are to be understood as operationally defined and discussed here.

than any other variable. Perhaps like penal or military institutions, learning can be at least partially satisfying and effective in dominated, oppressed groups as long as everyone is treated equally. It may be that when one inmate, rookie, or student is unfairly favored or set above the other, the energies of the group are diverted from the attainment of institutional or private goals into the resulting dissention.

Another group of structural measures that predict learning have to do with "organization" of the class—goal direction, disorganization, and formality. This group calls to mind Ryans' (1960) "Teacher Characteristics Pattern Y"—responsible, businesslike, systematic teacher behavior. A previous study (Walberg, 1968a) showed that these climate variables can be predicted from teacher personality. Among the organization measures, there are eight correlations with learning criteria.

Affective climate predictors can be grouped into "syntality" and "synergism" measures. One might derive from political theory the hypothesis that, like nationalism which promoted modern states, "syntality," or emotional identification with a group cause, enhances learning. Such does not appear to be true of the *class*, however; with one exception, the "syntality" measures—group status, classroom intimacy, and alienation—do not predict the criteria. Waller's hypothesis (1932) that students identify with the *school* through competitive *extramural* sports, pep rallies, social clubs, and the like may prove more fruitful empirically.

The climate measures of "synergism," the personal (or what some psychologists have termed the "psychodynamic" or "interpersonal") relations between class members, do predict learning. These variables are personal intimacy, friction, and satisfaction, and they account for 12 correlations with the criteria. Thus, it is not the identification with the group that correlates with learning but the perception that the class is personally gratifying and without hostilities among the members.

SUMMARY AND CONCLUSIONS

This is one of a series of exploratory studies derived from a sociopsychological theory of the classroom as a social system (Getzels & Thelen, 1960). In a national, nonrandom sample of 76 high school physics classes, it tested the hypothesis that individual perceptions of 18 structural and affective aspects of classroom climate predict 9 cognitive, affective, and behavioral learning measures adjusted for initial differences. Simple and multiple correlation revealed significant and complex relations between climate measures and learning criteria. For example, stratification and friction predicted science understanding, but other climate variables predicted physics achievement and attitudes toward laboratory work.

In addition, groups of climate variables predicted learning better than others. Among the structural variables, "isomorphism" (the tendency for class members to be treated equally; see Discussion for further explanation) and "organization" (efficient direction of activity) predicted learning much better than "coaction" (compulsive restraint or coercion). Among the affective climate measures, "synergism" (personal relations among class members) predicted learning better than "syntality" (identification with group goals).

Replications of the entire series of studies are being carried out with revised and, hopefully, more reliable instruments using a national random sample. Should the results hold up in other samples, especially in other school subjects, they should increase understanding of the social psychology of the class. Moreover, from a practical point of view, the ability to predict learning outcomes from assessments of classroom climate may have implications for teacher education, behavior modification of in-service teachers, and the assessment of teaching effectiveness, provided educators can agree on measurable goals of education.

REFERENCES

BEREITER, C. "Multivariate analyses of the behavior and structure of groups and organiza-

tions." In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology*. Chicago: Rand McNally, 1966.

COOLEY, W. W., & REED H. B. The measurement of science interests: An operational and multidimensional approach. *Science Education,* 1961, **45,** 320–326.

FERGUSON, G. A. *Statistical analysis in psychology and education*. New York: McGraw-Hill, 1959.

GAGE, N. L. (Ed.) *Handbook of research on teaching*. Chicago: Rand McNally, 1963.

GETZELS, J. W., & THELEN, H. A. The classroom as a unique social system. *National Society for the Study of Education Yearbook,* 1960, **59,** 53–81.

GUILFORD, J. P. *Psychometric methods*. New York: McGraw-Hill, 1954.

HARRIS, C. W. (Ed.) *Problems in measuring change*. Madison: University of Wisconsin Press, 1963.

OSGOOD, C. E., SUCI, G. J., & TANNENBAUM, P. *The measurement of meaning*. Urbana: University of Illinois Press, 1957.

ROBINSON, W. S. Ecological correlations and the behavior of individuals. *American Sociological Review,* 1950, **15,** 351–357.

RYANS, D. G. *Characteristics of Teachers*. Washington, D.C.: American Council on Education, 1960.

WALBERG, H. J. *Classroom Climate Inventory*. Cambridge: Harvard Project Physics, 1966.

WALBERG, H. J. Dimensions of science interests of boys and girls studying physics. *Science Education,* 1967, **27,** 111–116.

WALBERG, H. J. Teacher personality and classroom climate. *Psychology in the Schools,* 1968, **5,** 63–67. (a)

WALBERG, H. J. Structural and affective aspects of classroom climate. *Psychology in the Schools,* 1968, **5,** 247–253. (b)

WALBERG, H. J., & ANDERSON, G. J. The achievement-creativity dimension and classroom climate. *Journal of Creative Behavior,* 1968, in press.

WALBERG, H. J., & WELCH, W. W. A new use of randomization in experimental curriculum evaluation. *School Review,* 1967, **75,** 369–377.

WALLER, W. *The sociology of teaching*. New York: Wiley, 1932.

WELCH, W. W., & PELLA, M. O. The development of an instrument for inventorying knowledge of the scientific processes of science. *Journal of Research in Science Teaching,* 1967, **5,** 64–68.

(Received November 16, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 6, 420-424

WORD RECOGNITION BY CHILDREN OF TWO AGE LEVELS[1]

JAMES W. HALL
Northwestern University

2 word lists were presented aurally to kindergarteners and third graders. List 1 was presented under free learning instructions with half the *S*s at each age pronouncing words aloud after presentation. For List 2, *S*s reported whether each word had occurred in List 1. List 2 included 10 "new" words which were associates of List-1 words and 10 which were not (experimental—EX and control—C words, respectively). 5 List-1 words were repeated in List 2. More EX than C words were falsely recognized as having occurred in List 1, indicating that EX words occurred as implicit associative responses (IARs) during presentation of List 1. IAR-produced false recognitions were more frequent for younger than for older *S*s and for pronouncing compared with nonpronouncing *S*s. Recognition of repeated words was facilitated by overt pronouncing for kindergarteners.

There is evidence that when a single, familiar word is presented to a human *S*, at least two types of implicit responses may occur. One is the response involved in the act of perceiving the word, and has been termed the "representational response" (RR) by Bousfield, Whitmarsh, and Danick (1958). A second response has been called the "implicit associative response" (IAR) by Underwood (1965) and may be conceived as a second word elicited as an internal response by the stimulus properties of the RR. For example, if SCISSORS is presented, *cut* may occur as an IAR. Underwood has demonstrated that the occurrence of an IAR in this fashion may lead an *S* to erroneously identify the word that occurred as an IAR (*cut* in this case) as having been presented.

In a more recent study, it was found that IAR-produced false recognitions occurred in young children but were more frequent in five- and six-year-olds than in eight- and nine-year-olds (Hall & Ware, 1968). This age difference was contrary to expectations based on the notion that the older children would be more likely to produce IARs, thus more likely to become confused between the word presented to them and the IAR to that word.

One purpose of the experiment reported here is simply to replicate the age-difference finding of the Hall and Ware study. A second major purpose is related to a hypothesis regarding this age difference. More specifically, it is proposed that the older children make fewer IAR-produced false recognitions than do the younger ones, not because the older children produce fewer IARs, but because their ability to discriminate between RRs and IARs is greater than that of the younger children. If this is the case, then anything that affects discriminability of RRs and IARs should influence frequency of IAR-produced false recognition. One such variable may be the overt pronunciation of the presented word. Thus, in this experiment approximately half of the *S*s at each level were required to say aloud each word presented to them, and the remainder were given no instructions regarding pronunciation of the words.

METHOD

Subjects

The *S*s were 40 kindergarten children (17 boys and 23 girls) with mean chronological age of 5 years, 10 months and 40 third graders (20 boys and 20 girls) with mean chronological age of 8 years, 11 months, enrolled in a public elementary school in Winnetka, Illinois.

[1] This study was supported in part by a grant from the United States Office of Education. The author is grateful to Paul J. Avery, formerly Superintendent of Schools, and to Willis C. Mortensen, Principal of Crow Island Elementary School, in Winnetka, for their cooperation in providing *S*s and facilities.

TABLE 1

LIST-1 (FREE-LEARNING) WORDS USED AND THEIR FUNCTIONS

Word	Function	Word	Function
at	F	salt	CS
chair	F	eagle	R
money	F	girl	R
slide	F	gallop	CS
king	F	bed	CS
salt	CS	baby	R
eagle	R	thirsty	CS
gallop	CS	eating	CS
thirsty	CS	clear	R
girl	R	scissors	CS
scissors	CS	spider	CS
bed	CS	fingers	CS
baby	R	pretty	R
eating	CS	blossom	CS
pretty	R	lamp	CS
spider	CS	mouth	F
lamp	CS	pencil	F
clear	R	look	F
fingers	CS	window	F
blossom	CS	coat	F

Note.—Abbreviated: F = filler, CS = critical stimulus, R = repeated.

Design

The design called for the presentation of one word list under free learning (FL) instructions, followed by a second list under recognition instructions. List 1 is shown in Table 1; List 2, in Table 2. The words are listed in the order in which they were presented, with the function of each word indicated beside it. Included in List 1 were 10 critical stimulus (CS) words, each of which has been shown to elicit a particular response with relatively high frequency when standard word-association procedures are used. These high-frequency responses to the CS words, presumed likely to occur as IARs, were placed in List 2 as experimental (EX) words. In recent word-association data (Palermo & Jenkins, 1966; Entwisle, 1966), the mean frequency with which the 10 EX words used here were elicited by their respective CS words was 43.7% for kindergarteners[2] and 45.4% for third graders.

Also included in List 1 were five repeated (R) words, so termed because they also appeared in List 2. To reduce learning differences within List 1 due to serial position, 10 filler words occupied the first and last five positions in List 1. Each CS word and each R word appeared twice within List 1.

List 2 contained the 10 EX words, 10 Control (C) words, the five R words, and six new filler words. The C words were similar to the EX words in general frequency of occurrence (Thorndike & Lorge, 1944) but were not strong associates of any List-1 words. Thus, if false recognitions were more frequent for EX than for C words, it would be inferred that this difference was due to the previous occurrence of EX words as IARs.

Procedure

The procedure for each *S* (run individually) consisted of aural presentation of List 1 at a 5-second rate. Then, after a 7-minute interval, the List-2 items were presented at a 4-second rate. The *S*s were instructed to respond "yes" if a word had occurred on List 1 and "no" if it had not. The 7-minute delay between learning and recognition was used to increase the difficulty of the recognition task. To prevent rehearsal during that interval, *S*s were occupied with jigsaw puzzle tasks. All instructions and words were presented by use of a tape recorder.

During FL, an experimental variation in instructions was introduced. Twenty-four of the younger and 23 of the older *S*s, selected randomly, were instructed to *pronounce* each word aloud after it had been presented, and to attempt to remember the word. The remaining 20 younger *S*s and 23 older *S*s were instructed identically except that no request for pronunciation was made. All *S*s followed these instructions properly. For purposes of analyses, random procedures were used to exclude 4 *S*s from one group and 3 from each of two others, equalizing the *S*s at 20 per group.

TABLE 2

LIST-2 (RECOGNITION) WORDS USED AND THEIR FUNCTIONS

Word	Function	Word	Function
girl	R	house	F
needle	F	sleep	EX
train	F	ball	C
pepper	EX	gold	C
coffee	C	food	EX
lion	C	clear	R
horse	EX	web	EX
eagle	R	run	C
hair	F	baby	R
lazy	F	church	C
tall	C	hand	EX
cut	EX	flower	EX
water	EX	south	C
car	C	light	EX
pretty	R	receive	C
read	F		

Note.—Abbreviated. R = repeated, F = filler EX = experimental, C = control.

[2] Because the Palermo and Jenkins norms do not include data on kindergarteners, frequency estimates for 7 of the 10 critical response-experimental pairs were based on first-grade data.

RESULTS

False Recognitions

In Table 3 the mean numbers of false recognitions per S of the EX and C words are shown separately for each age level and each pronunciation condition. Using a difference score (EX − C) for each S, a t test for correlated data showed the overall mean of the differences ($\bar{X} = .81$, $SD = 1.30$) to be highly reliable, $t = 5.56$, $df = 79$, $p < .001$. This is interpreted as confirming earlier results by Underwood (1965), Davis (1967), and others in showing that EX words frequently are elicited as IARs during learning, resulting in their subsequent false recognition.

The EX − C difference scores then were used to examine the effects of age and pronunciation instructions on frequency of IAR-produced false recognitions. Analysis of variance showed the main effect of age to be highly reliable, $F = 7.48$, $df = 1/76$, $p < .01$. That is, the frequency of IAR-produced false recognitions was higher for the younger than for the older Ss. The main effect of instructions to pronounce also was highly reliable, $F = 10.89$, $df = 1/76$, $p < .01$, indicating that overt pronunciation reduced the frequency of IAR-produced false recognitions. The interaction between these variables was not significant.

TABLE 3

FALSE RECOGNITION OF EXPERIMENTAL (EX) AND CONTROL (C) WORDS AND CORRECT RECOGNITIONS OF REPEATED (R) WORDS

Instruction	Kindergarten			Third grade		
	Ex words	C words	R words	Ex words	C words	R words
Overt pronunciation						
$\bar{X}$ per S	.95	.35	4.60	.50	.35	4.60
SD	.76	1.37	.75	1.05	.59	.60
No overt pronunciation						
$\bar{X}$ per S	2.00	.25	3.65	1.25	.50	4.35
SD	1.69	.55	.74	1.57	1.09	.84

Correct Recognitions

The mean numbers of correct recognitions per S of the R words also are shown in Table 3. Since the data were characterized by marked skewness and nonhomogeneity, nonparametric analyses were performed in which groups were compared in terms of the number of Ss who correctly recognized all five R words. On this basis, the two third-grade groups and the kindergarten Ss under overt pronouncing instructions were quite similar. The mean numbers of perfect scores in these three groups were 13, 11, and 15, respectively. Of the nonpronouncing kindergarten Ss, however, only 3 of the 20 correctly recognized all five R words. For the kindergarten Ss, the difference in this respect between the two pronouncing conditions was highly reliable, $\chi^2 = 12.22$, $df = 1$, $p < .001$, while the corresponding difference between the two third-grade groups did not approach significance. Clearly, the instructions to pronounce did increase the frequency of correct recognitions of R words by the younger Ss. Although no similar effect was found for the third graders, it should be noted that the performance level of the nonpronouncing third graders was so close to maximum that the possibility of a ceiling effect must be considered.

DISCUSSION

False Recognitions

The fact that IAR-produced false recognitions were more frequent for the younger than for the older children simply replicates the earlier findings of Hall and Ware (1968). The question, then, is whether the older children simply produce fewer IARs under these experimental conditions or whether some other process is accounting for the differences found. The first alternative seems unlikely in view of evidence from other lines of research and common conceptions of verbal development in children. It is generally believed that implicit verbal behavior increases markedly from about 3–8 years of age. The Kendlers (e.g., 1963), in particular, have

provided considerable experimental evidence that verbal mediation increases substantially during this period. Presumably, IARs are frequently involved in such mediation, so that one would expect their occurrence to be more frequent in the older children, not less frequent. Of course, one might speculate that there is something about the particular experimental situation that inhibits the production of IARs by the older children, and that under other circumstances false recognitions would be greater for the older children. However, there is another alternative that appears more plausible, at least as a working hypothesis.

Perhaps as a child grows older (5–8 years, say) he does become more productive of IARs, but, at the same time, he also becomes better able to discriminate between words that were presented and words that he was reminded of. The basis for discrimination is not clear. One possibility involves the frequency hypothesis proposed by Ekstrand, Wallace, and Underwood (1966). This notion is that the RR occurs with greater frequency than does the IAR. That is, *S* probably says the presented word silently several times while an IAR is unlikely to be rehearsed. This frequency difference, then, may be the basis for *S*'s ability to respond correctly during recognition, that is, to successfully distinguish between the IAR and the RR. Possibly the older *S*s rehearse the presented words more than do the younger *S*s, so that the frequency discrepancy on which discrimination is based increases with age.

As predicted, pronunciation of the words during FL reduced false recognitions that were attributable to IAR occurrence. The interpretation favored by the author is that pronunciation of the words increased discriminability of those words from the IARs which they elicited, although the basis for increased discriminability is unclear. Again, frequency may be involved if it is assumed that the speaking of a word, at least for some *S*s, adds to the number of implicit rehearsals that the word receives, producing a higher frequency of occurrence than if the word had not been spoken. It also could be argued that the time taken to pronounce the word left less time for IAR production than was available for *S*s who did not pronounce the word. Thus, it simply may be that fewer IARs were made by *S*s who pronounced. These, as well as other possibilities, cannot be evaluated at present.

Correct Recognitions

In the correct-recognition data one finding stands out—the striking effects of instructions to pronounce on the performance of the younger children. The importance of pronouncing responses in a learning task of this type is well documented (e.g., Mechanic & D'Andrea, 1965). The present data suggest that with the younger children explicit instructions to pronounce markedly increase pronouncing and thus learning. It is not clear whether the pronunciation would need to be overt, as it was in the present instance, rather than covert in order to obtain this effect.

The fact that for the older *S*s only a slight difference occurred between the pronouncing and nonpronouncing groups may have been due simply to a ceiling effect. However, an alternative worth examining further is that by the time a child is 8- or 9-years-old, normal learning instructions produce covert pronouncing responses so regularly that explicit pronouncing instructions are superfluous.

REFERENCES

BOUSFIELD, W. A., WHITMARSH, G. A., & DANICK, J. J. Partial response identities in verbal generalization. *Phychological Reports,* 1958, **4,** 703–713.

DAVIS, G. A. Recognition memory for visually presented homophones. *Psychological Reports,* 1967, **20,** 227–233.

EKSTRAND, B. R., WALLACE, W. P., & UNDERWOOD, B. J. A frequency theory of verbal discrimination learning. *Psychological Review,* 1966, **73,** 566–578.

ENTWISLE, D. R. *Word associations of young children.* Baltimore, Md.: The Johns Hopkins Press, 1966.

HALL, J. W., & WARE, W. B. False recognition in children produced by implicit associative re-

sponses. *Journal of Experimental Child Psychology,* 1968, **6,** 52–60.

KENDLER, T. S. Development of mediating responses in children. *Monographs of the Society for Research in Child Development,* 1963, **28** (2), 33–52.

MECHANIC, A., & D'ANDREA, J. The role of articulation in response learning. In *Proceedings of the 73rd Annual Convention of the American Psychological Association.* Washington, D. C.: American Psychological Association, 1965.

PALERMO, D. S., & JENKINS, J. J. Oral word association norms for children in grades one through four. Research Bulletin No. 60, 1966, Department of Psychology, Pennsylvania State University.

THORNDIKE, E. L., & LORGE, I. *The teacher's word book of 30,000 words.* New York: Teachers College, Columbia University, 1944.

UNDERWOOD, B. J. False recognition produced by implicit verbal responses. *Journal of Experimental Psychology,* 1965, **70,** 122–129.

(Received November 17, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 6, 425-431

EFFECTIVENESS OF LEARNING FROM A PROGRAMMED TEXT COMPARED WITH A CONVENTIONAL TEXT COVERING THE SAME MATERIAL

WILLIAM J. DANIEL[1]
University of North Carolina at Chapel Hill

AND

PETER MURDOCH
University of Iowa

Studying a programmed textbook was compared with studying a conventional textbook to determine which method leads to better performance on a content examination. Both texts covered similar material on operant psychology. In a setting where 12 teaching assistants each taught 2 discussion sections with enrollments of about 22 students, 1 section studied the programmed text, and the other studied the conventional text. At the end of the semester, all students took an examination which contained several types of items. The results of the experiment favored the programmed text in that, for all 6 objective types of items and for 5 out of 8 essay items, the level of performance of the programmed text group was higher than the conventional text group ($p < .05$ or better from the analysis of variance).

One type of popular research on programmed instruction compares the effectiveness of learning from a particular programmed text or teaching machine with conventional material covering the same topic. The present study compares a programmed text on operant psychology (Holland & Skinner, 1961) and a comparable conventional text (Skinner, 1953) in an introductory psychology course. The relative effectiveness of learning from the two texts was measured by the final examination of the course.

Leith (1962) reviewed the literature on the effectiveness of programmed versus comparable regular material. He concluded that programmed texts save time over conventional texts but do not lead to better performance on a content examination. One of the few consistent findings in the research on programmed materials is that learning, as measured by an examination, is about the same for the programmed and conventional materials. However, the failure to demonstrate greater effectiveness for the programmed over the conventional materials may be due to methodological weaknesses in the previous research (Evans, 1965; Holland, 1965). Among the methodological problems, the following have been suggested: programs too short; gross dependent measures; small sample size; order and transfer effects not controlled; differences in information, procedures, instructions, and institutions not controlled when comparing the two types of material; criterion test too easy; and allowing *S*s to take home the materials (Paulus, 1966). In a study which avoided these weaknesses, Ripple (1963) compared the effectiveness of four groups taught the same material, where two groups were taught with programs and two were taught by some more "conventional" means. Tests of retention were given immediately after study of the material and 10 days later. The scores on the first test by both program groups were significantly higher than those of the conventional-method groups, but on the second test this difference was maintained for only one of the program groups. Thus, when methodological weaknesses are avoided, material which is programmed

[1] The late William J. Daniel initiated this research but died before the data analyses were completed or any of this report was written. The second author wishes to thank: Robert Hall for typing and mimeographing the examination, classifying the items into their types, organizing the scoring of the examinations, and helping on part of a data analysis; Darrel Bock for providing statistical advice without which this report could not have been written; Robert Callahan and Roger Wells for critically reading an earlier draft of the paper; and the instructors, Carolyn Cardall, John De Lorge, John Delk, Amy Kimura, Thomas Moore, Elizabeth Rose, William Rouse, Richard Sanders, Richard Sprott, Jane Webb, and Roger Wells.

may be learned better than the same material presented in conventional format.

The experiment reported here was performed during a semester of a college sophomore-level psychology course, with laboratory and discussion sections taught by graduate teaching assistants. The students were told they were research participants only after they had taken the final examination of the course. Because each instructor taught two sections, it was possible to assign the programmed text to one of his sections and the regular text to his other section. All students took the same final examination, which consisted of several types of items. A multivariate analysis-of-variance computer program (Bock, 1965, 1966) was used to assess the effects of the experimental variables on the several types of items.

METHOD

Subjects

The *Ss* were 577 students[2] enrolled in an introductory psychology course at the University of North Carolina at Chapel Hill.

Procedure

At course registration, the students were automatically assigned to 1 of 24 laboratory and discussion sections which met once each week for 2 hours during the semester. Five female and seven male graduate-student instructors each taught two of these sections, with enrollments ranging from 18 to 26 undergraduate students. The book by Holland and Skinner (1961) was randomly assigned as a required text to one section of each instructor, and the book by Skinner (1953) was assigned to his other section. The students in a particular section used only the text assigned to them.[3] The material covered in the sections was partly left up to the teaching assistant, provided he did the following: gave instruction on methodology in psychology and elementary statistics, demonstrated research to the class, organized the class to perform research, and tested the students by at least one midterm examination covering the above material and about half of the assigned textbook. The instructors met about once a week with the senior author to agree upon the work to be covered in class.

At the end of the course, all the students were tested on operant psychology by means of a 100-item examination. The test items came from a larger pool of items contributed by the teaching assistants, four of whom selected the items for the final examination. The first 92 items were objective, but there were 6 different types of objective items. The last 8 items were of the essay type.

The objective item types were the following:

Multiple-choice format. In this type of format, with a stem and four alternatives, the student picked out the one correct alternative. (*a*) Multiple-choice–A (MCA) tested knowledge of specific content (25 items). The MCA items were based on the more specific details of operant psychology as treated in the texts, such as how to read a cumulative record or the differences among the basic schedules of reinforcement. (*b*) Multiple-choice–B (MCB) tested knowledge of concepts and principles (24 items). The MCB items were based on the more general concepts and principles covered in the texts, such as the analysis of emotion or punishment. (*c*) Multiple-choice–C (MCC) tested response to new material (10 items). The MCC items required the students to apply the operant psychology approach to examples not fully covered in the texts, such as instilling certain behaviors or describing novel response differentiations. (*d*) Multiple-choice–D (MCD) tested for applications to everyday life (11 items). The MCD items required the students to apply the operant psychology approach to everyday situations not fully discussed in the texts, such as driving a car or working for wages.

Free-recall format. In this type of format the student filled in the blank space(s) in a sentence with one or two words. (*a*) Free-recall–C (FRC) tested response to new material (11 items). (*b*) Free-recall–D (FRD) tested for applications to everyday life (11 items). The bases of the FRC and FRD items were, respectively, like those of the MCC and MCD items.

The differences among the item types within the multiple-choice format and within the free-recall format were difficult to maintain. This was due to the fact that all teaching assistants wrote items without a clear knowledge of the differences among the types. It was felt, nevertheless, that the categories were sufficiently established to treat the examination as providing six objective item type scores for each *S*. The different kinds of objective items were randomly ordered on the test.

The essays were to be answered in about half a page of writing per question, such as discussing

[2] Two students were suspected of cheating on the examination, and their responses were discarded, leaving 575 *Ss* for the data analyses.

[3] After their examination, the students were requested to complete a 15-minute questionnaire. They were to indicate which text(s) they had studied, how much time they devoted to their text(s) relative to others, and how they liked their text(s) relative to others. These data will be presented in a supplementary report. Some students indicated that they had studied both the programmed and conventional texts. The data from these *Ss* were included in all analyses.

the Skinnerian analysis of the value of money or the development and treatment of fears.

The students' responses to the objective and essay items were scored by their instructors. The former were scored with the same punch-key answer sheets. Before the essay material was graded, the instructors met and developed common scoring criteria. Without knowing the student's name, the instructors used a 4-point scale (0–3) to evaluate each essay written by their students.

Data Analyses

The mean number of correct items, per section, on the six types of objective items (total N = 144) were analyzed in an analysis-of-variance design with three factors. In addition to overall multivariate F tests, univariate and step-down F tests were performed for each dependent measure for the sex-of-instructor effect, the between-instructor-within-sex effect, and the textbook effect. Textbook-by-instructor-within-sex was used as the error term in testing the significance of these effects.[4] Before analysis, the order of the dependent variables was specified in terms of their likelihood of accounting for the textbook effect. Based on previous research, the order was MCA, MCB, MCC, FRC, FRD, and MCD.[5]

The mean scores per section for each of the eight essays were analyzed in a manner similar to the objective item types. However, there was no specified order of importance for the essay items, so they were ordered as on the examination. In addition, the MCA and MCB objective item types were added as the first two dependent measures of the analysis (total $N = 240$).

Subsequent to the above analyses, reanalyses were performed on the objective items data and the essay items data.

The objective items were scored differently for the essay items analyses than for the objective item types analyses. In the latter analyses partial credit was given for some answers. However, this occurred on only a few items, so partial credits tended to make the within cell distributions less normal than without them. On the analyses of the essay items all partial credits were rounded off. This accounts for the slight differences in the scores of the same two objective item types in the analyses of the objective item types and of the essay items.

Results

The mean score for each section for each type of objective item and for each essay item were calculated and analyzed as indicated.

The results of the multivariate F tests for the equality of the mean vectors for the six objective item types were as follows. There was no evidence of a sex effect (p = .300). The between-instructors-within-sex effect was statistically significant (p = .010), indicating that the sensitivity of the experiment was increased by having each instructor teach a section which used the programmed text and a section which used the regular text. The textbook effect was also statistically significant (p = .002), which indicates that, as a set, the objective measures differed according to the kind of textbook studied.

The nature of the textbook effect is presented in Table 1. For each of the six objective item types, the mean number of correct items for the programmed text group minus the mean number of correct items for the conventional text group is positive. The sizes of the mean differences and the standard errors of the contrasts are given in Table 1. The six univariate F ratios for the textbook factor were all statistically significant ($p < .05$ or better). Thus, the univariate tests supported the hypothesis that test performance of the programmed text group is better than that of the conventional text group.

The step-down F tests indicated that when the effects of MCA (knowledge of specific content) were eliminated statistically from the other five dependent measures, the differences in test performance on the objective item types between the two text groups were no longer statistically significant (all p's > .05, Table 1). One interpretation of the step-down F ratios is that knowledge of specific content is learned better from a programmed text than a conventional text. Furthermore, this superiority transfers to enable the student who studied the programmed text to do

[4] The reader unfamiliar with the multivariate and univariate analysis-of-variance approach used here is referred to Bock (1965, 1966). The second paper uses the present study as "An excellent example of a *multivariate* study using the principle of blocking by teacher . . . [p. 834]." Unfortunately, the data supplied by Robert Hall to Darrell Bock for his analysis were mislabeled and contained an error in data computation; there are also typographical errors in two of the F ratios reported by him. This accounts for the differences between Bock's results and those reported here.

[5] R. D. Bock, personal communication, July 29, 1967.

TABLE 1

COMPARISON OF EXAMINATION PERFORMANCE OF THE PROGRAMMED VERSUS THE CONVENTIONAL TEXT GROUP ON SIX OBJECTIVE ITEM TYPES

Source	Item type					
	MCA	MCB	MCC	FRC	FRD	MCD
Text contrast (programmed–conventional text)						
Mean difference	1.70	1.63	.36	1.79	1.06	.86
Standard error of difference	.21	.35	.16	.25	.15	.08
Mean square[a]	17.33	15.90	.76	19.28	6.74	4.43
Univariate F	67.18**	21.60**	4.85*	50.98**	50.94**	106.12**
Step-down F	67.18**	.00	.12	4.20	3.50	.34

Note.—Abbreviated: MC = multiple-choice format; A = knowledge of specific content; B = knowledge of concepts and principles; C = responding to new material; FR = free-recall format; D = application to everyday life.

[a] $df = 1/11$.

* $p < .05$.

** $p < .0005$.

better on the more abstract types of objective items. For this interpretation to hold it must be shown that there are actual differences among the several objective item types relative to the textbook effect. On the other hand, if MCD, the most abstract of the objective item types, acts just like MCA in the multivariate F test, the examination is largely homogeneous relative to the textbook effect. These contrasting views were tested in an additional analysis of the objective items, where the order of the dependent variables was changed from MCA, MCB, MCC, FRC, FRD, and MCD in the first analysis to MCD, FRD, MCC, FRC, MCB, and MCA. The only values which can be different in the two analyses are the step-down F ratios. Table 2 gives the step-down F values for the two orders of the objective item types. Knowledge of specific content and applications to everyday life are both able to account for the textbook differences, and, therefore, the objective item types were largely homogeneous.

TABLE 2

STEP-DOWN F RATIOS FOR TWO ORDERS OF SIX OBJECTIVE ITEM TYPES FOR THE PROGRAMMED VERSUS THE CONVENTIONAL TEXT CONTRAST

Order 1	MCA	MCB	MCC	FRC	FRD	MCD
Step-down F	67.18*	.00	.12	4.20	3.50	.34
Order 2	MCD	FRD	MCC	FRC	MCB	MCA
Step-down F	106.12*	2.69	.87	.62	.01	.53

Note.—Abbreviated: MC = multiple-choice format; A = knowledge of specific content; B = knowledge of concepts and principles; C = responding to new material; FR = free-recall format; D = application to everyday life.

* $p < .0001$.

The analyses of the eight essay items were similar to those performed on the objective item types. In the first analysis, MCA and MCB were the first two dependent measures, while in the second analysis they were replaced by MCD and FRD. In the two analyses the values which can differ are the multivariate F ratios for the tests of the equality of the mean vectors, the step-down F ratios, and the univariate F ratios for the objective item types.

The results of both multivariate F tests for the equality of the mean vectors for the essay items were similar to those for the six objective item types. There was (*a*) no statistically significant sex effect ($p = .422$ in the first analysis and $p = .280$ in the second analysis) and (*b*) a statistically significant between-instructors-within-sex effect ($p = .001$ for both analyses). How-

TABLE 3

COMPARISON OF EXAMINATION PERFORMANCE OF THE PROGRAMMED VERSUS THE CONVENTIONAL TEXT GROUP ON TWO OBJECTIVE ITEM TYPES AND EIGHT ESSAY ITEMS

Source	Objective		Essay							
	MCA	MCB	1	2	3	4	5	6	7	8
Text contrast (programmed–conventional text)										
Mean difference	1.63	1.70	.29	.37	.31	.18	−.20	.24	.40	.17
Standard error of difference	.35	.21	.07	.08	.05	.07	.54	.11	.08	.08
Mean square[a]	17.33	15.90	.49	.81	.58	.20	.25	.33	.94	.17
Univariate F	67.18***	21.60**	14.74**	18.67**	35.54***	7.23*	.14	4.66	21.95**	4.49
Step-down F	67.18***	.00	.65	3.96	5.68*	.05	.06	1.90	2.59	.00

Note.—Abbreviated: MC = multiple-choice format; A = knowledge of specific content; B = knowledge of concepts and principles.
[a] $df = 1/11$.
* $p < .05$.
** $p < .005$.
*** $p < .0001$.

ever, the textbook effect was short of significance in the first analysis ($p = .083$), while it was significant in the second analysis ($p = .027$).

The nature of the textbook effect for the essay items is presented in Tables 3 and 4. As can be seen from Table 3, which presents the textbook contrasts, the hypothesis of greater learning from the programmed text than the conventional text is supported by the univariate F tests. For seven of the eight essays, the mean score for the conventional text is positive, and five of the relevant F ratios were statistically significant ($p < .05$ or better).

Table 4 presents the step-down F ratios for the two analyses of the essay items. When the effects of MCA are statistically removed from the remaining nine dependent measures, only for essay Number 3 does the difference in test performance between the two text groups remain statistically significant ($p < .05$). When the effects of MCD are statistically removed from the other nine measures, the textbook contrasts are all statistically nonsignificant ($p > .05$). Taken together these two analyses suggest that, for the essay items, knowledge of specific content and appli-

TABLE 4

STEP-DOWN F RATIOS FOR ANALYSES WITH FOUR OBJECTIVE ITEM TYPES AND THE EIGHT ESSAY ITEMS FOR THE PROGRAMMED VERSUS THE CONVENTIONAL TEXT CONTRAST

Order 1	Objective		Essay							
	MCA	MCB	1	2	3	4	5	6	7	8
Step-down F	67.18**	.00	.65	3.96	5.68*	.05	.06	1.90	2.59	.00
Order 2	Objective		Essay							
	MCD	FRD	1	2	3	4	5	6	7	8
Step-down F	113.45**	4.29	1.77	.03	3.16	.09	.18	1.50	2.48	2.97

Note.—Abbreviated: MC = multiple-choice format; A = knowledge of specific content; B = knowledge of concepts and principles; D = application to everyday life; FR = free-recall format.
* $p < .05$.
** $p < .0001$.

cations to everday life are not both able to account for the textbook differences.

DISCUSSION

Consideration of the univariate F tests supported the hypothesis of greater effectiveness of learning from a programmed text over a conventional text covering the same matter. Of the 14 contrasts of test performance of the programmed minus the conventional text group, 13 were positive and statistically significant ($p < .10$ or better). Nine of the F ratios relating to these differences were significant at $p < .01$ or better. For the one item on which the contrast was opposite to prediction, the difference was statistically nonsignificant ($p > .10$).

The objective test items were similar in format to the programmed material. It could thus be argued that the greater effectiveness of learning from the programmed over the conventional text was a function of transfer between text and test. However, since the essay items were similar in format to the conventional material, use of the transfer argument leads to the prediction that on the essays, the conventional text group will perform better on the examination than the programmed text group. This clearly was not the case.

Consideration of the step-down F analyses suggested that, relative to the treatment effects, the objective item types did not differ among themselves. The essay items appeared to differ from the MCA objective item type but not from the MCD objective type. The essay items were intended to be the most abstract type of item, but as it turned out, they were comparable to the MCD objective item type. Because the test item types were largely homogeneous, and because there was no clear à priori theoretical basis for differentiating among them, it was not possible to establish differential effects of the item types relative to the textbook differences. Further research is necessary with a data collection instrument designed to test hypotheses about the differential effectiveness of the item types (for example, it was initially expected that knowledge of specific content is learned better from a programmed text than a conventional text and that this relative superiority transfers to more abstract item types).

In research of the present type there is a problem concerning the extent to which the programmed and conventional texts are comparable. That is, it is not known to what extent they cover the same content, are equally difficult, are equally interesting, etc. The textbooks used here were written by the same author for the same purposes, so the books should be comparable in style, difficulty, and content, etc. The instructors were very familiar with both books and were asked to judge the comparability of the books by means of a suitable questionnaire. In addition, they indicated whether the examination items tended to be based more on one of the texts than on the other. According to the ratings of the instructors, the texts were comparable, and the examination did not favor one of the texts. It was felt, therefore, that the present experiment provided a fair comparison of the practical usefulness of the two texts by Skinner. The programmed text was more effective for teaching operant psychology than the conventional text. More specifically, compared with the conventional group, the program group on the average obtained a 10% higher score on each multiple-choice item type and a 7% higher score on each essay.

Lumsdaine (1961) has emphasized the need for more theoretically oriented research on programmed materials. While the present study presents little theory, it does provide a sophisticated approach for discovering what Lumsdaine calls absolute and contingent generalizations. An example of an absolute generalization is that test performance on all types of items is increased more by using the programmed text than the conventional text. Contingent generalizations express the predicted effects of one factor in relation to others with which it is expected to interact (Paulus, 1966). The present analysis-of-variance approach provides a method for obtaining empirical support for both kinds of generalizations. The use of overall mul-

tivariate effects and similarities among univariate effects leads to the establishment of absolute generalizations, while the use of specific univariate and step-down effects leads to the establishment of contingent generalizations.

REFERENCES

Bock, R. D. A computer program for univariate and multivariate analysis of variance. In, *Proceedings of the IBM scientific computing symposium on statistics*. White Plains, New York: IBM Processing Division, 1965.

Bock, R. D. Contributions of multivariate experimental designs to educational research. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology*. Chicago: Rand McNally, 1966.

Evans, J. L. Programming in mathematics and logic. In R. Glaser (Ed.), *Teaching machines and programmed learning, II*. Washington, D. C.: National Education Association, 1965.

Holland, J. G. Research on programming variables. In R. Glaser (Ed.), *Teaching machines and programmed learning, II*. Washington, D. C.: National Education Association, 1965.

Holland, J. G., & Skinner, B. F. *The analysis of behavior*. New York: McGraw-Hill, 1961.

Leith, G. Teaching by machinery: A review of research. *Educational Research*, 1962, **5**, 187–199.

Lumsdaine, A. A. Some theoretical and practical problems in programmed instruction. In J. E. Coulson (Ed.), *Programmed learning and computer-based instruction*. New York: Wiley, 1961.

Paulus, P. Review of research on the effectiveness of a programmed text compared with a conventional text covering the same matter. Unpublished paper (available from author), 1966.

Ripple, R. E. Comparison of the effectiveness of a programmed text with three other methods of presentation. *Psychological Reports*, 1963, **12**, 227–237.

Skinner, B. F. *Science and human behavior*. New York: Macmillan, 1953.

(Received November 18, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 6, 432–437

RELATIONSHIP BETWEEN FORMAL INTRALIST SIMILARITY AND THE VON RESTORFF EFFECT[1]

S. JAY SAMUELS
Department of Educational Psychology, University of Minnesota

Experiment I: 60 1st graders were randomly assigned to either a high, medium, or low stimulus similarity paired-associate list. During learning trials 1 stimulus in each list was printed in red and the other stimuli were in black. Transfer trials were given after each 5 learning trials. During learning trials in the high-similarity list more ($p < .001$) correct responses were given to the stimulus in red. At transfer there was a reversal of significance, and fewer correct responses ($p < .02$) were given to the stimulus formerly in red. On medium- and low-similarity lists the differences in correct responses to red and black stimuli were not significant at learning or at transfer. Experiment II: 30 college *S*s learned a high stimulus similarity paired-associate list. During learning trials 1 stimulus was in red. The other stimuli were in black. On transfer trials all stimuli were in black. Learning and transfer trials were alternated. The results were similar to those found in Experiment I for the high-similarity list. It appears that magnitude of the von Restorff effect is influenced by stimulus similarity. Failure to find positive transfer for the stimulus formerly in red was discussed in terms of attentional factors in learning.

Since Hedwig von Restorff first performed her landmark experiment, in which she demonstrated the more rapid learning of an item in a list which was different from other items in the list, much subsequent research has confirmed her original findings. The facilitation in learning a particular item in a list produced by isolating that item or making it distinctive from other items has come to be called "the von Restorff effect."

Knowledge of factors which influence the von Restorff effect is of theoretical and practical significance. Of theoretical importance is the effect of formal intralist similarity on the magnitude of the von Restorff effect. Other questions of importance relate to transfer and attentional factors in learning lists containing isolated items. An educator might argue that a technique which facilitates learning but which produces negative transfer is of spurious educational value. In fact, if during learning *S* attends to an aspect of the stimulus which will be removed during transfer, then one can predict that any superiority gained during learning would probably be lost at transfer. While questions of stimulus similarity, transfer, and attention are of theoretical interest, they are relevant to practical problems in education. For example, several reading primers published in Europe print new words and parts of words in distinctive colors. At the University of Pittsburgh a research program in the teaching of reading has each vowel phoneme associated with a color, and the graphemic representation of the vowel is printed in that color (Kjeldergaard & Frankenstein, 1967). While these innovations may facilitate learning, it is important to consider the effect at transfer when the incidental cues are removed.

One way to isolate in an experimental list is to print an item in a color different from the other items. Newman (see Wallace, 1965) found that when either the stimulus (S) or response (R) term in a paired-associate (PA) list was isolated by color, learning the pair was facilitated.

[1] Appreciation is extended to Joel Best for his contributions in Experiment I, and to Jim Palmer for his help in Experiment II.

This research was supported from grants to the University of Minnesota, Center for Research in Human Learning, from the National Science Foundation, the National Institute of Child Health and Human Development, and from the Graduate School of the University of Minnesota.

Another way to isolate is to insert an item of high meaningfulness in a list of low-meaningfulness items. Rosen, Richardson, and Saltz (1962) reported that learning was facilitated to a greater extent when a high-meaningfulness item was inserted in a low-meaningfulness list than when it was placed among items of high meaningfulness. They accounted for their results by noting that in a low-meaningfulness list, in which discriminability among items is poor, isolation serves to differentiate the items.

Another variable which should influence the von Restorff effect in a somewhat similar manner as does meaningfulness is formal stimulus similarity. Formal similarity refers to the degree of visual distinctiveness among items. The purpose of Experiment I was to determine the effect stimulus similarity and stimulus isolation have on the von Restorff effect when only a single stimulus within the list is isolated by color.

It was predicted that when stimulus similarity is high, rate of PA learning will be faster for the S-R pair in which the stimulus term is isolated than for S-R pairs in which the stimulus terms are not isolated; when stimulus similarity is moderate or low, there will be no difference in rate of PA learning between an S-R pair having an isolated stimulus term and S-R pairs having nonisolated stimulus terms. During learning trials an isolated stimulus term was presented in red, while the nonisolated stimulus terms were presented in black. After each five learning trials, a transfer test was given to determine if *S*s were able to give the correct response when a formerly isolated (red) stimulus term was presented in black.

EXPERIMENT I

Method

Subjects. Sixty first-grade *S*s with no reading problems were used. Twenty *S*s were randomly assigned to PA List 4-L, 20 to List 6-L, and 20 to List 8-L. The *S*s were then randomly assigned to rows in the design.

Design. A 4 × 4 repeated-measures Latin-square design was used, one for each of the three PA lists. As seen in Figure 1, the first column in the design contained the words which were isolated by printing them in red. Columns 2, 3, and 4 contained the nonisolated words which were printed in black. Each row in the design contained a different word which was isolated in red. The order of presentation for the words in that row was randomized for each trial. It is important to note that each *S* was required to learn responses for isolated and nonisolated stimuli for the row in the design to which he was assigned. Thus, this design permitted comparisons to be made of PA learning rate as well as for transfer of isolated and nonisolated words presented to the same *S*.

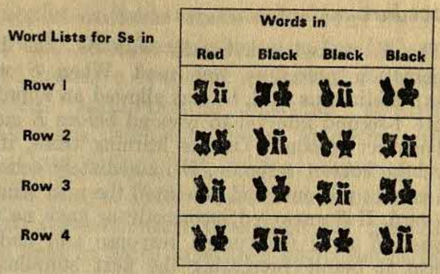


FIG. 1. Paradigm of 4 × 4 repeated-measures Latin-square design. (Each *S* learns PAs for words in row to which he is assigned.

Paired-Associate Lists

An artificial alphabet was used that had as little resemblance to English letters as possible. From the alphabet three lists of two-letter words were constructed (see Figure 2). List 4-L (high stimulus similarity) had four two-letter words constructed from only four different letters. List 6-L (medium stimulus similarity) had four two-letter words constructed from six different letters. List 8-L (low stimulus similarity) had four two-letter words constructed from eight different letters. The same response words were used for the three lists. They were: toy, dog, cat, man.

For the learning trials, three of the words in each list were printed in black, while the fourth was isolated by printing it in red. For the transfer trials, all the stimuli were printed in black. The stimuli were printed on 5 × 8 inch index cards.

FIG. 2. Stimuli and responses for Groups 4-L, 6-L, and 8-L.

Procedure

The *E* worked individually with *S*s. The PA anticipation procedure was used. When *S* was shown a stimulus card, he was allowed an approximate 4-second interval to respond before *E* gave the correct response. On the learning trials, if *S* gave the correct response, *E* immediately echoed the correct response and presented the next stimulus card. If *S* answered incorrectly or gave no response, *E* gave the correct response and had *S* repeat it before presenting the next stimulus.

After each five learning trials, a transfer test was given. During the transfer tests feedback was not given. The *S* was given a total of 15 learning trials and three transfer tests during the course of the experiment. For each of the three lists the stimuli were presented in the same random orders on the learning and transfer-test trials.

Results and Discussion

Table 1 shows the means and standard deviations for Lists 4-L, 6-L, and 8-L on learning trials and transfer tests. The 1-R in Table 1 refers to the words isolated in red in Column 1, while 2-B, 3-B, and 4-B refer to the nonisolated words printed in black in Columns 2, 3, and 4. On the transfer tests all the words were printed in black.

Learning. Analysis of variance for repeated-measures Latin-square designs were computed on correct responses given during learning for Lists 4-L, 6-L, and 8-L. The treatment effects of red versus black words were significant in List 4-L ($F = 10.87$, $df = 3/48$, $p < .001$), not significant in 6-L ($F = 1.75$, $df = 3/48$, *ns*), and not significant in List 8-L ($F = 1.22$, $df = 3/48$, *ns*).

In order to determine for the learning trials if the mean number of correct responses given to the isolated words was significantly different from the mean number of correct responses given to the nonisolated words, planned orthogonal comparisons were computed for Lists 4-L, 6-L, and 8-L. On high stimulus similarity List 4-L the differences between the means for isolated and nonisolated words were significant ($F = 30.79$, $df = 1/48$, $p < .001$). On medium stimulus similarity List 6-L the differences between the means for isolated and nonisolated words were not significant ($F = 3.76$, $df = 1/48$, *ns*). On low stimulus similarity List 8-L the differences between the means of the isolated and nonisolated words were not significant ($F = 3.06$, $df = 1/48$, *ns*).

Transfer. While significant differences during learning in number of correct responses were found in one of the three lists favoring the isolated over the nonisolated words, the critical issue was one of transfer or ability to give the correct response when the word formerly in red was presented in black. To compare number of correct responses for words which were presented in red during learning and in black during transfer with words presented always in black, *t* tests for correlated

TABLE 1
Means and Standard Deviations for Correct Responses during Learning and Transfer Reported by Columns within Similarity Groups

Item	Stimulus similarity											
	4-L				6-L				8-L			
	Columns				Columns				Columns			
	1-R	2-B	3-B	4-B	1-R	2-B	3-B	4-B	1-R	2-B	3-B	4-B
Learning												
M	9.85	5.90	4.70	5.75	10.75	9.80	9.50	8.70	11.90	10.00	10.45	10.80
SD	4.73	3.13	2.98	3.58	3.19	3.24	3.79	3.86	3.39	3.32	3.55	3.74
Transfer												
M	.50	1.05	.90	1.15	2.10	2.40	2.15	2.10	2.30	2.55	2.45	2.50
SD	.89	1.05	.79	.88	.85	.88	1.04	1.12	1.08	.69	.89	.83

Note.—R = words in red, B = words in black.

samples were computed. On high stimulus similarity List 4-L the pooled mean for number of correct responses given to the nonisolated words was significantly greater than the mean number given to the formerly isolated words ($t = -2.67$, $df = 19$, $p < .02$, two-tailed). On medium stimulus similarity List 6-L the pooled mean for number of correct responses given to the nonisolated words was not significantly different from the mean for the formerly isolated words $t \leq 1$, $df = 19$, *ns*). The same results were found for low stimulus similarity List 4-L ($t \leq 1$, $df = 19$, *ns*).

It would appear that in high similarity List 4-L, during learning trials, where discrimination on the basis of letter form was difficult, *S*s responded to isolated words on the basis of color. On transfer tests the color cue was absent, and *S*s did poorly with the formerly isolated words. On Lists 6-L and 8-L, where discrimination on the basis of letter form was easier, color was a less potent cue. For these lists it appears that during learning *S*s tended to use letter form as a cue for responding even for isolated words. One may find support for this interpretation by noting that on transfer tests for Lists 6-L and 8-L the means for isolated and nonisolated words were not significantly different from each other.

EXPERIMENT II

The failure in Experiment I to find positive transfer for the S-R pair in which the stimulus term was printed in red may be attributed to several sources. In the first experiment only three transfer tests were given. It is conceivable that *S*s did not have sufficient experience with the transfer task to realize that color was an irrelevant cue and letter shape the primary cue. Consequently, a pilot study was run with first graders using only a high-similarity list. The procedure was changed so that every learning trial was followed by a transfer test. Although this procedure provided ample opportunity for *S* to realize that for transfer tests color was an irrelevant dimension, the results were the same as in Experiment I (i.e., facilitation for the isolated pair during learning and negative transfer on tests where color cues were removed). Another explanation for failure to find positive transfer is that the children lacked sophisticated learning strategies and were unable to focus attention upon the less salient but critical dimension of letter shape. It is possible that college *S*s, with sophisticated strategies for learning, are able to focus attention upon the critical dimension of letter shape in the presence of the more salient dimension of color.

The purpose of Experiment II was to determine whether it would be possible to facilitate PA learning and transfer by (*a*) isolating with color one of the stimulus terms, (*b*) providing enough transfer tests for *S* to realize that in order to respond correctly at transfer he would have to focus on letter shape during learning, and (*c*) using *S*s with sophisticated learning strategies.

Method

Subjects. Thirty *S*s enrolled in introductory educational psychology were used.

Design. A repeated-measures design was employed in which each *S* learned one PA list containing S-R terms representing two conditions. In one condition during learning trials both terms of the S-R pair were printed in black. In the second condition during learning trials the S term of a pair was printed in red, while the R term was printed in black. On tests of transfer, S terms of both conditions were printed in black.

Materials. A high stimulus similarity list was used. The S-R pairs were: XYZ-*money*, XAF-*jewel*, XAZ-*kitchen*, VAF-*dinner*, ZYX-*office*, VUF-*garment*, ZOF-*wagon*, ZOV-*village*, VYX-*heaven*, FVY-*insect*. The list during learning trials contained nine S-R pairs which were printed in black letters, while in the tenth pair the S term was isolated by printing it in red and the R term was printed in black. For transfer-test trials all 10 S terms were printed in black, and no R terms were shown. Ten different PA lists were made. In each list a different S term was isolated in red. The lists were presented with a Lafayette memory drum.

Procedure

Three *S*s were assigned to each of the 10 PA lists according to order of appearance. The *S*s were run individually. Learning and transfer trials were alternated so that each *S* received seven learning and seven transfer tests. For the learning trials standard PA anticipation procedure was used. During the anticipation interval the S terms were exposed for 2 seconds followed by

exposure of the S-R terms together for an additional 2 seconds.

For the transfer tests each S term was presented alone for 4 seconds. No R-term feedback was provided during transfer tests. An 8-second intertrial interval was used between learning and transfer trials. The *S* was told that the same nonsense syllables would appear in different orders for learning and transfer, and he was to give the correct response associated with the stimulus as soon as he was able.

Results

Correct responses to the stimulus in red were compared to correct responses to the stimuli in black. Since there was but one stimulus in red and nine stimuli in black, proportions of correct responses to red and black over seven trials were calculated. Two matched-pairs *t* tests were computed, one for learning and one for transfer.

During learning the mean proportion of correct responses to stimuli in red was .59 ($SD = .33$) and to black, .32 ($SD = .13$). This difference was significant ($t = 4.23$, $df = 29, p < .001$, one-tailed).

During transfer the mean proportion of correct responses given to the stimuli printed always in black was .42 ($SD = .15$) and to the stimuli printed in red during learning and black during transfer, .27 ($SD = .32$). This difference was significant ($t = -3.41$, $df = 29$, $p < .002$, two-tailed).

DISCUSSION

Experiment II was conducted to determine if in a high stimulus similarity PA list the learning and transfer of an S-R pair could be facilitated by isolating with color the S term of the pair. The procedure employed alternating learning and transfer trials and using *S*s with sophisticated learning strategies. The results disclosed that during learning trials, when color cues were present, significantly more correct responses were given to the isolated than to the nonisolated S terms. During transfer, when color cues were removed, significantly more correct responses were given to the nonisolated S terms. Correct responses to isolated S terms on learning trials and failure to do so at transfer indicated that *S*s were attending to stimulus color and not stimulus shape. Although color was a relevant dimension during learning trials, it was an irrelevant dimension during transfer. The fact that *S*s focused primarily on the irrelevant dimension of color rather than the relevant dimension of shape for the isolated S-R pair is surprising first, because the number of alternations between learning and transfer trials was sufficient to indicate that color cues were being removed at transfer, and second, because *S*s were told that the same nonsense syllables would appear at learning and at transfer. It thus appears that when college *S*s learn a high stimulus similarity PA list in which an S term has been isolated by printing it in color they find it difficult to focus on a less salient, but critical, cue in the presence of a more salient cue. In this respect the learning strategy of college *S*s was similar to that of the children in Experiment I.

Several conclusions may be drawn from the findings. The von Restorff effect seems to be a less universal phenomenon than previously thought; that is, there are conditions under which increasing the distinctiveness of an item in a list does not facilitate learning. The von Restorff effect in PA learning seems to be reliably found in lists of high stimulus similarity where discriminability among items is poor. By isolating a stimulus in a high-similarity list, that item becomes distinctive in contrast to the other smtiuli, and the isolated pair is learned more rapidly than the nonisolated pairs. The von Restorff effect is not reliably produced in low stimulus similarity PA lists where the items are already highly discriminable from each other. A qualifying factor to this conclusion may be list length. If a low-similarity list is long and difficult to learn because of its length, it is possible that in this list isolation may facilitate learning. Thus, stimulus similarity, isolation, and list length may interact.

Another finding of interest relates to the difficulty *S*s had in shifting attention from a more salient but irrelevant cue to a less salient but relevant cue. A similar finding was reported in a study on attentional processes in reading (Samuels, 1967). In

this study children had to learn PAs. During learning trials the stimulus contained both a picture which could reliably elicit the correct response and letters which spelled the word. On transfer trials only letter stimuli were presented. Although learning and transfer trials were alternated, on learning trials *S*s attended to the cue which most reliably elicited the correct response (i.e., pictures) and consequently did poorly on transfer tests when picture cues were omitted. The conclusion one may reach is that in verbal learning, when the stimulus complex has several dimensions upon which *S* may focus attention, the principle of least effort operates (Underwood, 1963), and *S* tends to focus upon that dimension of the stimulus which most reliably elicits a response leading to reinforcement.

The tendency for *S*s to focus upon the stimulus dimension which most reliably elicits the correct response, even though the stimulus dimension may be irrelevant in terms of transfer, suggests that educational innovations which attempt to facilitate learning by using incidental cues must be evaluated in terms of transfer when the incidental cues are removed.

REFERENCES

Kjeldergaard, P. M., & Frankenstein, R. *Grapheme-phoneme regularity and its effects on early reading: A pilot study.* Pittsburgh: University of Pittsburgh, Learning Research and Development Center, 1967.

Rosen, H., Richardson, D. H., & Saltz, E. Meaningfulness as a differentiation variable in the von Restorff effect. *Journal of Experimental Psychology,* 1962, **64**, 327-328.

Samuels, S. J. Attentional process in reading: The effect of pictures on the acquisition of reading responses. *Journal of Educational Psychology,* 1967, **58**, 337-342.

Underwood, B. J. Stimulus selection in verbal learning. In C. N. Cofer & B. S. Musgrave (Eds.), *Verbal behavior and learning: Problems and processes.* New York: McGraw-Hill, 1963.

Wallace, W. P. Review of S. E. Newman. Paired-associate learning as a function of stimulus term and response term isolation. In, Review of historical, empirical, and theoretical status of the von Restorff phenomenon. *Psychological Bulletin,* 1965, **63**, 410-424.

(Received December 1, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 6, 438–448

STRATEGY SELECTION AND INFORMATION PROCESSING IN HUMAN DISCRIMINATION LEARNING[1]

L. BERELL KORNREICH
University of Wisconsin, Milwaukee

In 2 experiments, adult human *S*s received 32 4-dimensional discrimination problems. A method which purported to tap all the hypotheses (Hs) was compared with Levine's (1966) blank-trial procedure, which taps only 1 H. No difference in *S*'s behavior under these 2 procedures was found. Each *S*'s strategy was inferred from the pattern of Hs manifested, and it was found that *S*'s strategy contributed a large amount of the variance. Separate analyses for different strategies were performed. It was pointed out that whether outcomes are preprogrammed or contingent upon *S*'s choice response, the effect of outcome, that is, whether "right" or "wrong," is obscured by merely analyzing the problems solved. An analysis was performed which looked at information-processing errors after each outcome. It was found that most errors occur on Trial 2 of the individual problem and that significantly more errors are made after "wrong" than "right" outcomes on both Trials 1 and 2. However, the difference is not large considering the number of trials involved.

Levine (1966) introduced a theory of human discrimination learning which treated *S* as an information processor and analyzer. He proposed that *S*s operate under a strategy almost identical to Bruner's (Bruner, Goodnow, & Austin, 1956) "focusing" strategy. It is assumed that *S*s attempt to remember (encode) all the cues which logically could be correct after an outcome (e.g., "right" or "wrong"). These cues are then stored in memory as "hypotheses" (Hs) and are tested against future outcomes. In this way, Hs are eliminated from the retained set until one H remains as the "correct answer" to the problem.

A methodology was employed whereby the set of possible Hs was determined by the experiment, and one H that *S* was holding after each trial could be inferred. This inference permitted unambiguous prediction of the response on outcome trials after the first one. The nature of the outcome, whether right or wrong, was controlled so that the effect of outcome on the retention or rejection of the H being held could be analyzed.

Levine's procedure for inferring one H being held involves presenting four "blank" (no outcome) trials. There are 16 possible sequences of *S*'s four choice responses. The stimuli were constructed so that eight sequence patterns conform to the eight possible Hs as defined by *E* and told to *S*. Thus, if *S* keeps responding on the basis of a single H, that H manifests itself in a distinguishable sequence over the four blank trials.

Levine found that *S*s were very systematic in this respect, and that they continued to respond on the basis of an H until a wrong outcome was received. Then they chose another H. An analysis of the new Hs chosen revealed that *S*s had to be holding several Hs at one time, and that they also eliminated several at one time. This finding gave rise to the formulation of the "focusing" strategy.

The "blank-trial" procedure allowed for the inference of only one H, and the analysis of the grouped data revealed that *S*s held several Hs at one time. The analysis of grouped data further showed that the information-processing ability of *S*s was far from perfect. This was especially apparent after wrong outcomes. So the question arises: What causes the group to make errors? It could be that all *S*s use the same strategy, that errors occur due to faulty encoding at some point during the

[1] This article is based on a dissertation submitted to the Graduate School, Indiana University, in partial fulfillment of the requirements for the PhD degree. The author wishes to thank his advisor, Richard Young, for his help and encouragement throughout the dissertation project.

problem, and that *S*s differ in their ability to encode. Levine's theory makes such an assumption. It could also be that *S*s differ in the strategy used, so that some errors are due to inferior strategies, as well as faulty implementation of the focusing strategy. In any event, a procedure which taps all the Hs being considered by *S* would shed light on the question of how errors occur. It would also be a direct test to Levine's implicit assumption that all *S*s use the same or a similar strategy.

It was decided that a choice-response procedure would be interpolated between outcome trials instead of the four blank trials used by Levine. In the present experiment, *S*s were faced with eight buttons and asked to indicate which Hs "still could be correct" after each of the outcomes. It was hoped that such a procedure would tap all the Hs being considered by *S* after each outcome without radically changing the task used by Levine. In order to check on the possibility that the task had been altered, groups of *S*s were run under Levine's blank-trial procedure in the first experiment.

EXPERIMENT I[2]

Method

Subjects. The *S*s were 60 undergraduate volunteers at Central Connecticut College.

Stimulus cards and problems. The discrimination problems consisted of cards on which there were drawn two stimuli about 1½ inches apart. The stimuli varied in four dimensions—color, size, letter, and position. Black and white were used as colors, X and T, as letters. A large letter was 1 inch, a small letter, ½ inch in height. All the problems were composed of a series of such cards, with a blank card separating the problems.

Two groups of 20 *S*s received identical problems which incorporated Levine's blank-trial procedure. A problem was composed of 16 cards, 4 outcome cards and 12 blank-trial cards. Outcomes, either right or wrong were always given after Cards 1, 6, and 11 and were given half the time after Card 16. Figure 1 shows a diagram of the problem.

These four cards form a set with several properties. Each value of each dimension is combined exactly twice with the values of all the other dimensions. The set provides that, after the first outcome, four of the eight cues remain as logically possible solutions; after the second outcome, two remain; and, after the third outcome, the solution is logically determined. This is true whether the outcomes are right or wrong and allows *E* to program all outcomes for *S*. It also follows that S always has only a 50% chance of choosing the correct stimulus on the first three outcome cards. Such a set is also used as blank trials and insures that eight of the 16 possible sequences of choices over the four cards correspond to the eight cues or "Hs."

The four cards used as outcome cards are called Set A. After the first three outcome cards, four blank-trial cards were presented which were also orthogonal because they were merely the same stimuli reversed on the card. These four cards are called Set B. Thus, instead of the large-black-X being on the left, it appeared on the right, etc., for all four cards. The four blank trials were used in order to infer *S*'s H after the first three outcomes. The entire problem is presented in diagram form in Figure 2.

Design and procedure. The major comparison in this experiment is between the blank-trial method used by Levine to indicate one H being held by *S* after each outcome, and a method which requests *S* to indicate all Hs being held after each outcome. Therefore, a group run under each method was required. A third group was added run under Levine's method, but with the apparatus of the other method present, in order to assess the effect of having the eight possible Hs on display for *S*.

All three groups received the same outcome cards, and *S*s were assigned at random to one of the three groups. The *S*s run under Levine's blank-trial procedure were given 16 such problems; *S*s in Group 2 repeated these 16 problems with different outcomes and so received 32 problems in all. A deck of stimulus cards was prepared which made up eight problems. A random order was used for both outcome (Set A) and blank-trial (Set B) cards. The deck was merely turned back to the beginning, and the cards were presented again for the remaining problems. The *S*s were told that the same solutions would not necessarily be correct, even though the cards were being presented a second time. Orders of cards and the written order of the Hs on the apparatus were changed after five *S*s had been run in each group. Thus, four orders were used, since each group contained 20 *S*s.

Another important part of the design relates to the orthogonality of the stimulus-card sets. Levine programmed the outcomes, whether right or wrong, for each problem. This was possible since either outcome is "logical" until the fourth outcome. He found that no *S* indicated doubt that the problems had predetermined solutions.

Each sequence of three outcomes was presented twice over the 16 problems. That is, the sequence of right-wrong-wrong was presented twice, etc.,

[2] The author wishes to acknowledge the assistance of Frederick Karls in the collection of the data for Experiment I.

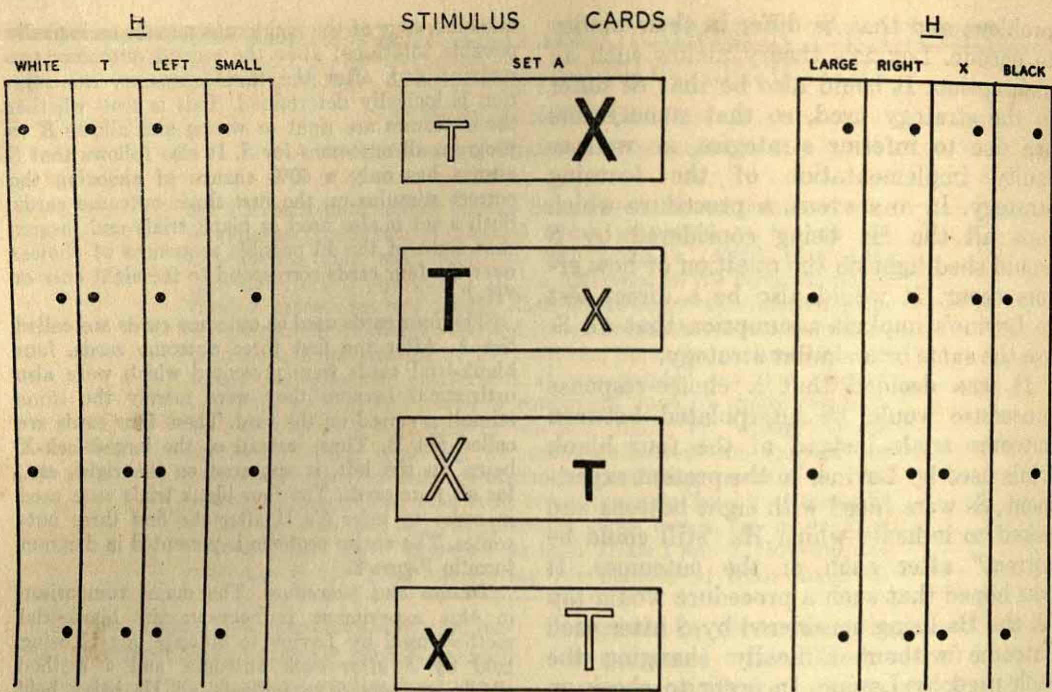


FIG. 1. A set of four stimulus cards and the eight sequence patterns produced when one hypothesis is followed by *S* over the four trials.

for the other seven sequences. This procedure was followed for all three groups. For Group 2 each sequence was again presented twice during the second 16 problems, only in a different order.

The programming of outcomes allowed for systematic testing of the differential functioning of right and wrong. Yet, a problem can still be scored as solved if the last H used by *S* is consistent with the "information" contained in the first three outcome trials. Thus, if *S* chose the black stimulus and was told "wrong" on the first three outcome trials, the white cue can be regarded as the correct solution. If *S* evidenced a white H after the third outcome trial, he can be said to have solved the problem.

Thus the design consisted of three groups of 20 *S*s each. Groups 1 and 1A were nearly identical, in that each was run under the blank-trial procedure. The only difference between them was that Group 1A had the wooden apparatus with the eight Hs written on it together with additional instructions commenting upon the presence of the apparatus.

Each *S* was instructed and run individually. The instructions used for the 40 blank-trial *S*s were nearly identical to those used by Levine. The only difference was due to giving *S*s only two instead of four practice problems. The first practice problem consisted of 10 trials in which the color (black) was the basis for solution. The *S* received the deck face up. He responded to the top card, the appropriate outcome was given, and he then turned the card face down out of the way. This procedure was followed throughout the experiment. The second practice problem consisted of 46 trials with an outcome given at the first trial and at every fifth trial thereafter. The left position was the basis for solution. In both practice problems, nonsolving *S*s were given corrective instructions, redid the problem, and continued on.

After completing the practice problems all *S*s received the same instructions. They were told that "the problems will all be like the ones you've just had, always with one of these simple solutions. That is, one of the colors, sizes, letters, or positions will be correct." They were informed that "a problem has just ended and a new one is beginning when you turn over a blank card."

Group 1A received the following additional instructions:

> The eight possible answers are on display there, so that you don't have to memorize them. Remember, only one of these eight possible answers is correct for each problem. It is your job to find out which one is correct.

Group 2 received the following additional instructions:

> If that is clear, I'd like to add one more feature. You will notice that this small box in front of you has eight buttons. Each button has one of the possible answers written above it. Remember, only one of these eight possible answers is correct for each problem; and it is

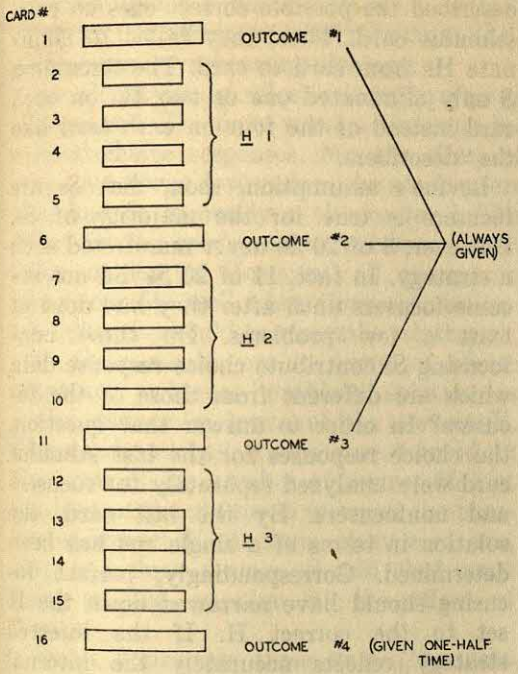


Fig. 2. A diagram of the 16-trial problem showing the outcome (right or wrong) trials and the blank (nothing said) trials from which *Ss'* hypotheses (Hs) were inferred.

consistently correct for that problem. It is your job to find out which one of these eight is the correct one for that problem. After I say "right" or "wrong," you'll press one or more buttons to indicate which of the answers you think still could be correct. Any questions about that?

The apparatus used for Group 2 was a completely mechanical wooden box (13 inches wide × 8 inches long). It had a panel with eight buttons on it which faced *S*. Above each button was written one of the eight Hs in the following manner: X, T, large, small, white, black, left, right. The two levels of one dimension were always kept together. When a button was depressed by *S* it tipped over a small wood block on a hinge inside the box. Since the back of the box was open, the tipped blocks were visible to *E* but not to *S*, and since the button returned to its former position when released, no visible record of the buttons pressed remained for *S*. The *E* kept a record of the buttons pressed. The blocks were then reset after each trial to the upright position.

Results and Discussion

The first question to be considered is whether the introduction of a visual representation of the eight Hs and the interpolation of a button-pushing task radically changed the task used for Group 1. In order to answer this question, the number of correct-choice responses to the last stimulus cards of the problems was compared among the three groups. Out of 320 problems, Groups 1, 1A, and 2 made 68%, 72%, and 75% correct-choice responses, respectively. An analysis of variance for these differences revealed no significance, $F = 1.19$, $df = 2/48$. Neither the effect of the order of Hs on display nor the interaction of order and groups was statistically significant. Since there is no significant difference between groups, it can be assumed that the problem-solving task is the same for all groups, and that a visual record of the eight Hs does not improve problem solving after instructions define the eight Hs for *S*. For the rest of the analyses, the data from Groups 1 and 1A (blank-trial groups) will be combined.

In this experiment, the stimuli were all black and white, and X and T. Levine used different colors and letters. A comparison of data from the blank-trial groups in this experiment with Levine's data reveals that the change in stimuli and *Ss* had not greatly changed *S* behavior. Levine found that 92.4% of the blank-trial patterns corresponded to an H, whereas the corresponding figure for this experiment is 89.9%. Levine found that when a right was given, the same H was retained 95% of the time (based on interpretable patterns). When a wrong was given, a different H was selected 98% of the time. The corresponding figures for Groups 1 and 1A are 91% and 96%, respectively. So, although the present data are slightly less systematic than Levine's, his type of "H" analysis is applicable.

The *Ss* in Group 2 produced two sets of data; one set indicated their choice responses on the four stimulus cards, and the other set indicated their choice of Hs after each outcome had been given. The purpose of having *Ss* indicate all the Hs they were holding after each outcome was to test Levine's (1966) theory. He proposed that an attempt is made to "focus" from trial to trial on the basis of the outcomes (information) given. For these problems, *Ss* who operate under a focusing strategy

should hold four Hs after the first outcome, two Hs after the second outcome, and one H after the third outcome. Other strategies may be manifested by a different pattern of responding.

An analysis of each *S*'s strategy was performed for the 20 Group 2 *S*s. A focusing strategy was operationally defined as 4-2-1 response pattern for three of four successive problems, which is the pattern such a strategy would be expected to produce. By this criterion, 14 of the 20 *S*s became focusers by the fifth problem; nine *S*s began as focusers. All 14 *S*s maintained that strategy for the rest of the 32 problems. These 14 *S*s will be treated as a group, and all 32 problems per *S* will be included in the remaining analyses.

Of the remaining six *S*s, three maintained nonfocusing strategies for the entire 32 problems. The other three adopted a focusing strategy at the twelfth, fourteenth, and twenty-third problem, respectively, and maintained a focusing strategy for the rest of the 32 problems. Thus, a "focusing strategy" may be conceived as an internal state that *S* may or may not enter. Entrance into this state seems to be an end state for *S*. A group of nonfocusers was formed from these six *S*s by including the first 16 problems of the three *S*s who eventually became focusers. The first 16 problems were included rather than just those problems before which they became focusers in order to preserve an equal number of types of problems. For example, there were two right-right-right and two wrong-wrong-wrong problems among the 16 problems.

Four of the six nonfocusers indicated four Hs after each trial. These four Hs represented the four cues of the inferred-correct stimulus on the just previous outcome card. That is, if they had chosen the small-black-X on the left and had been told "wrong," they would press four buttons indicating the large-white-T and right Hs, regardless of the outcome trial. Another *S* responded similarly except that he usually indicated only three of the four inferred-correct Hs. These *S*s could be called "describers," in that they merely described the possible correct cues on each stimulus card. Thus, they failed to eliminate Hs from card to card. The remaining *S* only eliminated one or two Hs on each card instead of the four on each card like the "describers."

Levine's assumption, then, that *S*s are focusers is true for the majority of *S*s. However, 3 of 20 *S*s never manifested such a strategy. In fact, 11 of 20 *S*s did not become focusers until after they had done at least a few problems. Do these nonfocusing *S*s contribute choice-response data which are different from those of the focusers? In order to answer that question, the choice responses for the last stimulus card were analyzed separately for focusers and nonfocusers. By the last card, the solution in terms of a single cue has been determined. Correspondingly, perfect focusing should have narrowed down the H set to the correct H. If the inferred strategy reflects accurately the internal process of *S*, nonfocusers should not have narrowed down the correct H and so should perform at an inferior level compared with focusers, in terms of making the correct-choice response on the last card.

In order to investigate whether the focusers corresponded to "good" problem solvers from Groups 1 and 1A (blank-trial groups) and whether nonfocusers corresponded to "bad" problem solvers, good and bad groups of *S*s were formed from Group 1 and 1A *S*s. "Good" *S*s were defined as those *S*s who manifested the correct H for at least 11 of the 16 problems. "Bad" *S*s were those who manifested only from three to seven correct Hs of the 16 problems. There were 13 good *S*s and 16 bad *S*s from Groups 1 and 1A.

Table 1 presents the data for the four newly created groups. The upper two rows compare the focusers and good *S*s. These data give strong support to the notion that good problem solvers under either a blank-trial or a button-pushing procedure are good precisely because they are using a focusing strategy. The lower two rows compare the bad *S*s and the nonfocusers. This comparison strongly supports the contention that some *S*s are poor problem solvers,

whether under a blank-trial or a button-pushing procedure, precisely because they do not employ a focusing strategy.

In all groups, an effect related to wrongs is apparent. The more wrongs, the more incorrect-choice responses, but the effect of even one wrong is to reduce the nonfocusers and bad *S*s to nearly a chance level. The effect for the good problem solvers is much more moderate. However, all groups do about equally well when the outcomes are all rights. It should be remembered, though, that these results are in terms of choice responses, whereas problem solution is really defined in terms of the correct H. The good and bad groups were formed on the basis of the number of correct Hs. An analysis of the groups in terms of the number of correct H solutions, given zero, one, two, and three wrongs, should shed more light on the question of the effect of wrongs.

Table 2 presents these data. The data for the good and bad *S*s are based only on the interpretable H patterns. The data for the focusers are based on those problems on which only one button was pushed after the third outcome. Since the nonfocusers always pressed more than one button, it was impossible to score their problems for the number of correct solutions in terms of the correct H. Again, the correspondence between the good *S*s and focusers is striking. The effect of wrongs for them is again to reduce the number of problem solutions moderately. The effect of wrongs for bad *S*s is drastic.

TABLE 1

Percentages of Correct-Choice Response for the Last Trial of the Problem for Four Specially Selected Groups of Subjects Given Zero, One, Two, or Three Wrongs in the Problem

	Wrongs			
	0	1	2	3
Good *S*s (from Groups	96	88	79	77
1 and 1A)[a]	26	78	78	26
Focusers (from Group	93	86	79	68
2)[b]	56	168	168	56
Bad *S*s (from Groups	91	58	51	53
1 and 1A)[c]	32	96	96	32
Nonfocusers (from	94	59	59	61
Group 2)[d]	18	54	54	18

Note.—The lower numbers indicate the n for that percentage, and n refers to the number of problems. Thus, the 26 means that there was a total of 26 right-right-right problems which were scored for the good *S*s. The response to the last card was correct for 96% of those 26 problems.

[a] 13 *S*s.
[b] 14 *S*s.
[c] 16 *S*s.
[d] 6 *S*s.

TABLE 2

Percentages of Problem Solution in Terms of Correct Hypotheses for Three Specially Selected Groups of Subjects Given Zero, One, Two, or Three Wrongs in the Problem

	Wrongs			
	0	1	2	3
Good *S*s (from Groups	100	86	75	72
1 and 1A)[a]	25	76	76	25
Focusers (from Group	91	87	78	65
2)[b]	54	154	148	52
Bad *S*s (from Groups	84	52	31	17
1 and 1A)[c]	31	79	81	29

Note.—The lower numbers indicate the n for that percentage, and n refers to the number of problems. It was not possible to determine the percentages for the 6 nonfocuser *S*s from Group 2.

[a] 13 *S*s.
[b] 14 *S*s.
[c] 16 *S*s.

Levine grouped all *S*s before he analyzed the data for the effect of right and wrong outcomes. From the data presented here, it can be seen how that procedure would produce a misleading estimate of the effect of wrongs. He found a striking effect for wrongs which may have been contributed mainly by nonfocusing *S*s. Instead, Levine assumed that a focusing strategy produced the effect of wrongs, and he proposed an explanation of the effect in terms of an assumed coding procedure used by focusers. It was assumed that *S* codes the cues of the stimulus he chooses. Therefore, when faced with a wrong, *S* has to "erase" the just previously coded information and recode the new information. For instance, consider that the Hs large-black-T-right

are still possibly correct after the first trial. The *S* chooses a large-right stimulus on the second trial and is told wrong. According to Levine, *S* must erase "large" and "right" and find "black" and "T." The advantage to *S* of receiving rights is that there is no necessity to recode.

Such an explanation, of course, does not speak to the data of the nonfocusing *S*s. Does such an explanation even apply to the data from focusers? Before an answer to that question can be given, an important artifact of the procedure must be discussed. In order to study the effect of right and wrong outcomes, Levine predetermined the schedule of outcomes for each problem. A problem still had a solution, but it was defined as a function of *S*'s choice responses and the outcomes given. For example, if *S* happened to choose the left side on all three outcome trials of a problem designated as a wrong-wrong-wrong problem, the solution became defined as "right" for that *S* on that problem. If *S* manifested a "right" H after the third outcome, he was credited with solution of the problem. Thus, *S* always defined the solution for himself. This procedure was used in the present experiment.

Although such a procedure looks like a clever way to control the outcome variable so as to better study its effect, just the opposite results. Consider an *S* who remembers only one of the four correct Hs after the first trial. If two rights occur on the second and third outcome trials, he will "solve" the problem merely by responding consistently on those two trials. The same *S*, if confronted with a wrong on the second trial, will have to abandon his one stored outcome. He can only infer that the correct H is one of the four represented by the other stimulus, the position of *S*s after the first trial. Thus, an *S* who has "learned" exactly the same thing as another, will not be credited with solving the problem because wrongs happened to occur in that problem. Similar contingencies exist for all the combinations of Hs held and outcomes.

The high percentage of right-right-right problems solved by bad *S*s can now be explained. These *S*s merely had to respond consistently on the basis of one correct H after the first trial. The rights would serve to define that H as the correct solution. The effect of right and wrong outcomes cannot be analyzed in terms of problems solved due to the artifact caused by predetermining the outcomes. Before an appropriate analysis is presented, Experiment II will be introduced. Then the analysis will use the combined data of Experiments I and II.

Experiment II

Experiment II *S*s were run under conditions similar to Group 2 *S*s in Experiment I. Again *S*s received 32 problems and were asked to indicate between choice trials which "answers still could be correct." However, a major change involved presenting "real" rather than "preprogrammed" outcomes. That is, a solution was selected for a problem (i.e., "black"), and outcomes were presented depending upon *S*'s choice response, rather than the program set by *E* as in Experiment I.

Method

Subjects. Thirty-four student-nurse volunteers who were taking their psychiatric training at Connecticut Valley Hospital, Middletown, Connecticut, served as *S*s.

Stimulus cards and problems. The problems were the same as those used in Experiment I except that real outcomes were presented. The same order of cues appeared on the wooden apparatus for all *S*s.

Procedure. The instructions were shortened by eliminating the two practice problems. A paragraph was added at the end which restricted the number of buttons pressed to only one after the third and fourth cards. This change in procedure was made so that the data for nonfocusing *S*s could be analyzed to determine what H they were following in their choice responses. Richter (1965) has reported choice-response data for bad *S*s which were below the .5 guessing probability for problems with two and three wrong outcomes. It was hoped that knowledge of the H such bad *S*s were using to make choice responses would help to explain their below-chance responding.

Results and Discussion

A group of bad *S*s was selected in terms of their choice responses on the fourth

trial. Twleve *S*s who made 13 or more errors constituted the group. The number of errors as a function of wrongs in the problem was computed. Below-chance responding was not observed for any problem type, so that below-chance responding in Richter's group may have been due to his selection procedure.

Only 9 of the 34 *S*s became focusers by the criterion of a 4-2-1 pattern of button responses for three of four problems in succession. All but one of these *S*s became a focuser before the ninth problem. Of the remaining *S*s, 11 were describers, and 7 were partial focusers, in that they either failed to hold all four Hs after the first trial or they failed to press four buttons. The other 7 *S*s, as indicated by a postexperimental interview and inspection of their data, were either seeking a sequence solution or only eliminated one H when a wrong was given.

Again it is clear that all *S*s are not focusers. The number of focusers in a group may be dependent on the extensiveness of the instructions and the number of practice problems or the population from which *S*s are drawn.

Can it then be said that nonfocusers "do not understand" the instructions? Such a question forces a consideration of what it would mean to understand. The avowed purpose of the instructions for Levine and Richter was to restrict the H set to the eight, and to establish that a solution consisted of a single H. If "understanding" means to have learned this, then the only *S*s who can be said not to have understood are the four *S*s who looked for a solution in terms of sequences. These *S*s held other than the eight Hs.

In a sense, however, any strategy other than a focusing strategy for these simple four-trial problems indicates a lack of understanding. Only focusing allows for a "solution" by the fourth trial in the sense that correct Hs arrived at by other strategies involve guessing or luck. The major "learning" which takes place, then, is the formation of a strategy. The *S*s can be viewed as being in one of two states; they either have learned or they have not learned with respect to the task set them.

General Discussion

The remaining analyses will consider only focusers, those *S*s who really solved the discrimination problems. The effect on information processing of right versus wrong outcomes can only be considered with respect to their data, since, for nonfocusers, successive trials do not carry information in a measurable sense. Table 3 shows a comparison between focusers from Experiments I and II. For the Experiment II group, only the problems after *S* became a focuser are included. The correspondence between the two groups is striking. Using real problems appears to have little effect. Again, there is a steady decrease in solutions as more wrongs serve as outcomes. However, before these data are taken as evidence for greater difficulty in information processing after wrongs, an artifact must again be considered.

In Experiment II, the correct cue was determined by the H, and the outcomes were allowed to vary as a function of the solution and *S*'s choice. Consider an *S* who forgets the correct cue after the first trial. The probability of receiving a right on the second trial is less for him than for an *S* who also forgot one H, an incorrect one. This is because the correct H is no longer in his set of Hs, whereas, for the other *S*, his set of Hs has been reduced by one incorrect H. On the third trial, one *S* will receive a wrong, since he is only holding one H, an incorrect one, whereas the other *S* will at least have a 50% chance of receiving a right if he still holds two Hs, the correct one and an incorrect one. In the extreme case, an *S* who happens to remember only the correct cue by chance will receive all rights if he merely chooses on the basis of this H. In this way, an information-processing error will lead to wrongs and an incorrect solution if the error involves the cue designated correct for that problem by *E*. An information-processing error which is just as significant from a psychological point of view will increase the

TABLE 3

Percentages of Problem Solution in Terms of the Correct Hypothesis for Focusers from Experiments I and II Given Zero, One, Two, or Three Wrong in the Problem

Focusers	Wrongs 0	1	2	3
From Exp. I[a]	91	87	78	65
	54	154	148	52
From Exp. II[b]	97	87	74	64
	30	89	89	25

Note.—The lower numbers indicate the n for that percentage, and n refers to the number of problems. Thus, the 54 means that there was a total of 54 right-right-right problems which were scored for the focusers from Experiment I. The correct hypothesis was manifested after the third trial for 91% of these problems.

[a] 14 *S*s.

[b] 9 *S*s.

probability of rights and correct solution if the error does not involve the correct cue. For this reason, the data presented in Table 3 are not an appropriate measure of the effect of rights versus wrongs. Thus, whether one preprograms outcomes or allows them to vary, the effect of rights versus wrongs is obscured.

In both Experiments I and II, an appropriate measure involves observing the entire set of Hs manifested after right and after wrong outcomes. After Trial 1, four Hs may be considered correct, in that complete information processing would encode all four Hs embodied by the positive stimulus. After Trial 2, two Hs would be considered correct, and, after the third trial, one H remains as the solution. In order to analyze errors after the first trial, then, the Hs manifested by *S* are compared with the four correct Hs. An error may involve getting three out of four, etc. This type of error is called an error of omission. Another type of error would consist of *S*'s pressing five or more buttons. This is called an error of inclusion.

The analysis for Trials 2 and 3 is the same except that fewer Hs are correct. The results of this analysis are presented in Table 4. It should be noted that errors which occur on Trial 1 and are merely "carried" to Trial 2 do not count as errors on Trial 2. The same procedure also applies for Trial 3.

It can be seen that most errors involve forgetting one H of the correct set and either substituting an incorrect H for it or merely manifesting one less than the complete set. In considering only this type of error from both experiments, a Z test indicates a significant difference in errors after right and wrong outcomes. For Trial 1, $Z = 2.03$, $p < .05$, and for Trial 2, $Z = 2.78$, $p < .01$, which shows a greater probability of a -1 type error after wrong than after right for Trials 1 and 2.

Not every information-processing error results in an incorrect solution. In order to account for incorrect solutions, that is, an incorrect H after the third trial, a different analysis is needed. Only those problems on which incorrect solutions occurred are selected. Then each problem is evaluated to determine at what trial and whether after a right or wrong the correct cue was dropped from the held set. Table 5 presents these data for both Experiments I and II. The proportion of unsolved problems is .18 for Experiment I and .16 for Experiment II. A focusing strategy is associated with more than 80% solution of these simple problems. Both sets of data show the same pattern of errors. On Trial 2, information-processing errors account for over half of the incorrect solutions. Also on this trial wrong outcomes appear to cause more incorrect solutions than right outcomes. A Z test on the combined data of Experiments I and II for Trial 2 indicates a difference between right and wrong outcomes significant beyond the .01 level. However, this difference is a small one considering the total number of problems.

Levine assumed that a coding process of the sort proposed recently by several authors (Glanzer & Clark, 1964; Haber, 1964; Sperling, 1963) takes place. Levine further assumed that *S*s code, that is, attempt to remember, the cues represented by the stimulus they choose *before* an outcome is given. In this way, he attempted to account for what he took to be greater diffi-

TABLE 4

Number and Type of Information-Processing Errors after a Right (+) and after a Wrong (−) for Trials 1, 2, and 3

Trial	Errors of omission				Errors of inclusion		% errors on each trial
	−1	−2	−3	−4			
1[a]							
Exp. I	18+ 25−	1+ 9−	0+ 1−	1+ 4−	3+	8−	Exp. I $= \frac{69}{431} = .16$
Exp. II	2+ 10−	0+ 1−	0+ 0−	1+ 2−	0	1−	Exp. II $= \frac{17}{234} = .07$
2[b]							
Exp. I	26+ 49−	7+ 10−			17+	18−	Exp. I $= \frac{127}{431} = .29$
Exp. II	17+ 24−	1+ 6−			2+	2−	Exp. II $= \frac{52}{234} = .22$
3[c]							
Exp. I	11+ 16−				6+	3−	Exp. I $= \frac{36}{431} = .08$
Exp. II[d]	6+ 6−						Exp. II $= \frac{12}{234} = .05$

Note.—The −1, −2, −3 and −4 refer to an error of missing one, two, three, or all four of the correct hypotheses.

[a] Four correct hypotheses.

[b] Two correct hypotheses.

[c] One correct hypothesis.

[d] Errors of inclusion disallowed by procedure.

culty in coding after wrong outcomes. If *S* codes the stimulus he chooses on the first trial before an outcome is given, the encoded set remains the correct one if a right outcome is given, whereas the encoded set must be "erased" and its complement coded if a wrong outcome is presented.

The reported data show that a small but statistically significant difference in coding errors exists as a function of outcome on the first trial. The difference is explained by Levine's coding hypothesis. The reason for the small difference involves the procedure used by Levine and in the reported experiments. Both the positive and negative instances were presented simultaneously. After an outcome, *S* controlled the exposure time since each *S* turned the card over at his own rate. Therefore, after an outcome was presented, *S* could observe the inferred-correct stimulus for as long as he wished. It is proposed that *S*s quickly learn to do that. In other words, *S*s learn to encode the H set after and not before the outcome. Before they learn, wrongs do pose more of a problem than rights. After learning, when told "wrong" after pointing to a stimulus on the first trial, *S* merely observes the other stimulus on the card and codes those four Hs. Interviews with *S*s after the experiment support such an interpretation. Apparently *S*s can code the four Hs almost perfectly.

TABLE 5

Information-Processing Errors Which Resulted in Incorrect Solutions Presented by the Trial on Which They Occurred and Whether after a Right (+) or Wrong (−) Outcome for Experiments I and II

Trial	Exp. I	Total errors	Exp. II	Total errors
1	1+ 5−	6 = .08	1+ 2−	3 = .08
2	13+ 31−	44 = .57	5+ 17−	22 = .59
3	11+ 16−	27 = .35	6+ 6−	12 = .33
Total		$\frac{77}{422} = .18$		$\frac{37}{234} = .16$

Note.—In Experiment I nine problems could not be scored since two hypotheses were manifested on the third trial.

A reliable, although again not a large, difference in information-processing errors as a function of outcome was observed after the second trial. More errors are made after the second trial than either the first or third. An error is made after the second trial on about 25% of the problems, and it is these errors which lead to over half of the incorrect solutions. Why should this be true? The *S* has only to code two Hs rather than four, as after the first trial. The answer must lie with a perceptual factor. The *S* must remember the four correct Hs from Trial 1 as he chooses on Trial 2. If given a right, he must seek the two correct cues represented by the chosen stimulus. If given a wrong, he can either seek the two correct cues of the chosen stimulus and subtract them from the four encoded ones, or seek the two correct cues represented by the other stimulus on the card. The few more errors after wrongs is probably due to some *S*s attempting to make subtractions. The greater number of errors after the second trial is probably due to forgetting one of the four cues once the first card is turned over. There is also perceptual interference from the incorrect cues which are part of each stimulus. Few errors either after a right or wrong are made after Trial 3. Apparently, when only two Hs need be remembered, the lack of perceptual aids is not critical.

From the two studies reported here, it can be seen that a large source of variance in the data was accounted for by the strategy which *S* used. Therefore, the previous procedure of grouping the data from all *S*s together led to inaccurate estimates of important parameters. The failure to recognize the artifact introduced by preprogramming outcomes and by the guessing probability led to further inaccuracies. For these reasons, Levine (1966) and Richter (1965) emphasized the assertion that wrong outcomes cause more difficulty in information processing than right outcomes, rather than individual differences in *S*s' approach to the experimental task.

REFERENCES

BRUNER, J. S., GOODNOW, J. J., & AUSTIN, F. A. *A study of thinking*. New York: Wiley, 1956.

GLANZER, N., & CLARK, W. W. The verbal loop hypothesis: Conventional figures. *American Journal of Psychology,* 1964, **77,** 621–626.

HABER, R. N. Effects of coding strategy on perceptual memory. *Journal of Experimental Psychology,* 1964, **68,** 357–362.

LEVINE, M. Hypothesis behavior by humans during discrimination learning. *Journal of Experimental Psychology,* 1966, **71,** 331–338.

RICHTER, N. Memory, choice, and stimulus sequence in human discrimination learning. Unpublished doctoral dissertation, Indiana University, 1965.

SPERLING, G. A model for visual memory tasks. *Human Factors,* 1963, **5,** 19–31.

(Received December 28, 1967)

Journal of Educational Psychology
1968, Vol. 59, No. 6, 449–451

STORAGE AND RETRIEVAL OF WORDS IN FREE-RECALL LEARNING

JAMES H. CROUSE
State University of New York at Binghamton[1]

2 free-recall learning experiments were performed. In Experiment I the presence or absence of retrieval cues was varied factorially with the presence or absence of storage cues. Retrieval cues facilitated recall when storage cues were present, but not when they were absent. Experiment II showed that the facilitating effects of retrieval cues depended on the pairing of storage cues and the words to be recalled. The results were interpreted in terms of the storage tags generated during input.

A trial in a typical free-recall learning (FRL) experiment consists of two phases, the items being presented for study during the input or storage phase, and recall being tested during the output or retrieval phase. In a recent study, Tulving and Pearlstone (1966) presented nouns belonging to categories during the input phase of FRL. Category names of the nouns were always presented along with the nouns and served as storage cues (e.g., weapons—BOMB, CANNON; crimes—TREASON, THEFT). During the output phase the same category names were presented as retrieval cues in some conditions but not in others. Recall of the nouns was highest when retrieval cues were presented. This effect of retrieval cues may be dependent on the storage cues being presented during input. That is, the effect of retrieval cues may be greater when storage cues are presented than when they are not presented. This hypothesis is tested in Experiment I.

EXPERIMENT I

Method

Every *S* was given a single trial on the same list of 35 nouns. Each noun was taken from a different category in the Cohen, Bousfield, and Whitmarsh (1957) norms with the mean category frequency being 29.17. The design was a 2 × 2 factorial in which the first variable was the presence or absence of storage cues and the second was the presence or absence of retrieval cues. The four conditions formed by these variables are designated as C-C, C-NC, NC-C, and NC-NC. Twelve undergraduates from the State University of New York at Binghamton were assigned to each condition. The *S*s were run in small groups numbering from one to five. The groups were randomly assigned to the conditions subject to the restriction of achieving an equal number per condition.

The nouns were read at a 3-second rate. When storage cues were presented, *S*s were told that a category name (the category label from the Cohen et al., 1957, norms) would be read before each noun (e.g., a bird—RAVEN). They were told that they would not have to recall the category names, but that the category names would help them recall the nouns. When storage cues were absent, only the nouns were read (e.g., RAVEN). Immediately following presentation, a 4-minute recall period was given. When retrieval cues were presented, the 35 category names were listed in the same order as they occurred during input with a blank space beside each one. Brief instructions at the top of the recall sheet told *S* that each noun presented could be described by one of the categories and that he should print each noun he could recall beside its appropriate category. When retrieval cues were absent, only the blank spaces were provided, and brief instructions at the top of the recall sheet told *S* he should print the nouns in these spaces. The *S*s in both retrieval cue conditions were told to recall the nouns in any order they desired. They were not informed of their condition of recall until the time of recall.

Results

The number of nouns correctly recalled by each *S* was computed (Table 1). Recall was higher when storage cues were presented than when they were absent, F (1, 44) = 5.54, $p < .05$, and recall was higher when retrieval cues were present than when they were absent, F (1, 44) = 16.49, $p < .01$. The most important finding, however, was the significant interaction, F (1, 44) = 10.51, $p < .01$, which was further analyzed by individual F tests. It

[1] Now at the University of Delaware.

TABLE 1
MEAN NUMBER OF CORRECT RESPONSES

Storage cues	Retrieval cues	
	Absent (NC)	Present (C)
Absent (NC)	13.33	14.33
Present (C)	12.25	21.17

was found that Group C-C recalled significantly more nouns than each of the other groups, the *F*s (1, 44) being 26.67, 15.66, and 20.58, for the comparisons with Groups C-NC, NC-C, and NC-NC, respectively. None of the comparisons among the latter three groups approached significance, *F*s (1, 44) < 1.46.

Discussion

This significant interaction shows that retrieval cues produce greater recall than no retrieval cues when storage cues are presented, thus replicating the findings of Tulving and Pearlstone (1966), but have no effect on recall when storage cues are absent. Essentially this same finding also has been obtained by Wood (1967) in work reported after the completion of the present research. In Experiment I Wood reported a significant interaction in which retrieval cues had a larger effect when storage cues were presented, but, unlike the present experiment, the effect of retrieval cues was not eliminated completely when storage cues were absent. The effect was eliminated, however, in Experiment II when category frequency was comparable to the present study.

It seems possible, as was the case in Experiment I above, that the facilitation which occurs from retrieval cues when storage cues are presented (Tulving & Pearlstone, 1966; Wood, 1967) may be due to the appropriate pairing of the cues with the nouns during storage, but it may also be due to some other aspect of presenting the category names during storage. Experiment II determines this possibility by comparing three conditions: C-C, C_I-C, and C-NC. In Condition C_I-C category names are presented as storage and retrieval cues, but the cues are inappropriately paired with the nouns during storage. Recall should be greater in Condition C-C than in C-NC, again demonstrating the effectiveness of retrieval cues when storage cues are present and appropriately paired. If the appropriate pairing of storage cues and nouns is important for this effect of retrieval cues, then recall in Condition C-C should be greater than in C_I-C; however, if the effect of retrieval cues results from the presentation of the storage cues per se, then Condition C-C should not differ from C_I-C.

EXPERIMENT II

Method

Sixteen undergraduate female students from the State University of New York at Binghamton were assigned to each of the three conditions: C-C, C_I-C, and C-NC. The method was the same as Experiment I except for the following: (*a*) In the C_I-C condition the category names of the nouns (the category labels from the Cohen et al., 1957, norms) were randomly paired with the nouns as storage cues and subsequently presented at recall as retrieval cues; and (*b*) *S*s were told that the names to be presented might or might not help them recall the nouns. The same thing was added to the brief instructions at the top of the recall sheet for *S*s having retrieval cues.

Results

The mean number of nouns recalled was 23.94, 12.50, and 12.38 for Conditions C-C, C_I-C, and C-NC, respectively. An analysis of variance on these groups was significant, F (2, 45) $= 37.95$, $p < .01$. Individual F comparisons indicated that recall was higher in Condition C-C than Condition C-NC, F (1, 45) $= 57.53$, $p < .01$, thus showing the facilitation from retrieval cues when storage cues are present and appropriately paired. Recall was also higher in Condition C-C than Condition C_I-C, F (1, 45) $= 56.30$, $p < .01$, suggesting that the facilitation produced by retrieval cues in Condition C-C is associated with the appropriate pairing of the storage cues and nouns during presentation and not simply the presentation of the cues per se. Recall did not differ significantly in Conditions C_I-C and C-NC, $F < 1.0$.

Discussion

The presentation of an appropriate category name as a storage cue may lead to a stable association between the category name and a noun. That is, the storage cue may provide the category name as a storage tag for a noun (e.g., Yntema & Trask, 1963). When storage cues have been presented, the presentation of retrieval cues would reinstate the storage tags so that the associations could be used to retrieve the nouns, and recall would be higher than when the retrieval cues are not presented to reinstate the storage tags. However, when storage cues have not been presented, storage tags are not provided. While *S*s likely produce their own storage tags, it would seem unlikely that each noun is tagged by a category name. Therefore, the presence of retrieval cues would not lead to greater recall than the absence of retrieval cues. In summary, the principle that emerges is that retrieval cues may facilitate recall when they are successful in reinstating the storage tags that were formed during input.

REFERENCES

Cohen, B. H., Bousfield, W. A., & Whitmarsh, G. A. Cultural norms of verbal items in 43 categories. Technical Report No. 22, 1957, University of Connecticut, Contract Nonr-631, Office of Naval Research.

Tulving, E., & Pearlstone, Z. Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior*, 1966, **5**, 381–391.

Wood, G. Category names as cues for the recall of category instances. *Psychonomic Science*, 1967, **9**, 323–324.

Yntema, D. B., & Trask, F. P. Recall as a search process. *Journal of Verbal Learning and Verbal Behavior*, 1963, **2**, 65–74.

(Received December **21, 1967**)

Page 1

STATEMENT OF OWNERSHIP, MANAGEMENT AND CIRCULATION
(Act of October 23, 1962: Section 4369, Title 39, United States Code)

Publisher: File two copies of this form with your postmaster.
Postmaster: Complete verification on page 2

Form Approved, Budget Bureau No. 46-R029

1. DATE OF FILING: October 1, 1968

2. TITLE OF PUBLICATION: Journal of Educational Psychology

3. FREQUENCY OF ISSUE: Bi-monthly

4. LOCATION OF KNOWN OFFICE OF PUBLICATION *(Street, city, county, state, ZIP code)*: Mount Royal and Guilford Avenues Baltimore, Maryland 21202

5. LOCATION OF THE HEADQUARTERS OR GENERAL BUSINESS OFFICES OF THE PUBLISHERS *(Not printers)*: 1200 17th Street N.W. Washington, D.C. 20036

6. NAMES AND ADDRESSES OF PUBLISHER, EDITOR, AND MANAGING EDITOR

PUBLISHER *(Name and address)*: American Psychological Association, 1200 17th Street N.W., Washington, D.C. 20036

EDITOR *(Name and address)*: Wayne H. Holtzman University of Texas, Austin, Texas 78710

MANAGING EDITOR *(Name and address)*: Helen Orr 1200 17th Street N.W., Washington, D.C. 20036

7. OWNER *(If owned by a corporation, its name and address must be stated and also immediately thereunder the names and addresses of stockholders owning or holding 1 percent or more of total amount of stock. If not owned by a corporation, the names and addresses of the individual owners must be given. If owned by a partnership or other unincorporated firm, its name and address, as well as that of each individual must be given.)*

NAME	ADDRESS
American Psychological Association	1200 17th Street N.W., Washington, D.C. 20036

8. KNOWN BONDHOLDERS, MORTGAGEES, AND OTHER SECURITY HOLDERS OWNING OR HOLDING 1 PERCENT OR MORE OF TOTAL AMOUNT OF BONDS, MORTGAGES OR OTHER SECURITIES *(If there are none, so state)*

NAME	ADDRESS
None	

9. FOR COMPLETION BY NONPROFIT ORGANIZATIONS AUTHORIZED TO MAIL AT SPECIAL RATES *(Section 132.122, Postal Manual)*
(Check one)
The purpose, function, and nonprofit status of this organization and the exempt status for Federal income tax purposes: [X] Have not changed during preceding 12 months [ ] Have changed during preceding 12 months *(If changed, publisher must submit explanation of change with this statement.)*

10. EXTENT AND NATURE OF CIRCULATION	AVERAGE NO. COPIES EACH ISSUE DURING PRECEDING 12 MONTHS	ACTUAL NUMBER OF COPIES OF SINGLE ISSUE PUBLISHED NEAREST TO FILING DATE
A. TOTAL NO. COPIES PRINTED *(Net Press Run)*	9882	10011
B. PAID CIRCULATION 1. SALES THROUGH DEALERS AND CARRIERS, STREET VENDORS AND COUNTER SALES	NONE	NONE
2. MAIL SUBSCRIPTIONS	8449	8705
C. TOTAL PAID CIRCULATION	8449	8705
D. FREE DISTRIBUTION *(including samples)* BY MAIL, CARRIER OR OTHER MEANS	97	109
E. TOTAL DISTRIBUTION *(Sum of C and D)*	8546	8814
F. OFFICE USE, LEFT-OVER, UNACCOUNTED, SPOILED AFTER PRINTING	1336	1,197
G. TOTAL *(Sum of E & F—should equal net press run shown in A)*	9882	10011

I certify that the statements made by me above are correct and complete. *(Signature of editor, publisher, business manager, or owner)* Henry S. Salmon Jr Bus Mgr

POD Form 3526 May 1966

Journal of Educational Psychology
1968, Vol. 59, No. 6, 452–455

A CROSS COMPARISON OF WRITTEN AND SPOKEN RESPONSES WITH VISUAL AND AUDITORY CONFIRMATION IN A PAIRED-ASSOCIATE SIMULATED READING TASK

JANICE T. GIBSON AND BARBARA L. HAYDEN
University of Pittsburgh

64 college students learned a list of novel-symbol–nonsense-syllable paired associates in a task that simulated early reading. A 2 × 2 factorial design was employed in which *S*s either wrote or spoke their answers, and in which confirmation of results was either visual or auditory. All experimental groups required the same number of trials to reach criterion in the initial learning task. Later, the same groups decoded triphoneme "words" composed of the learned symbols. The *S*s who wrote their answers made fewer errors than those who spoke them. There was a significant interaction between response and confirmation modes in this task. With written responses, visual confirmation produced fewer errors. With spoken responses, auditory confirmation produced fewer errors.

The process of learning to read has never been explained in its entirety. As a result, methods of teaching reading often are haphazard, and attempts to correct reading problems often are unsuccessful. In order to develop more effective methods of teaching reading, it would be helpful to first analyze the reading process, break it down into its components, and then determine the variables affecting each component. This approach is not without precedent: Gibson (1965) discussed the value of analyzing beginning reading as a perceptual decoding process and experimentally manipulating the variables affecting it.

The reading process also can be considered a verbal learning task. Reading experts and other learning theorists already have compared the beginning stages of reading by phonics to paired-associate (PA) learning (Fries, 1963; Levin, Watson, & Feldman, 1964; Piekarz, 1963). In the PA task, *S* learns to make a specific and appropriate response each time a nonsense syllable is presented visually. Beginning readers also learn to make appropriate responses to what initially are nonsense graphic stimuli. A child learning by phonics often learns to decode a new word letter by letter without first having to hear the word in its entirety or to know what it means. This happens frequently when a phonic alphabet such as the Pitman Initial Teaching Alphabet is used.

Verbal learning tasks have been used in the laboratory to study some of the variables affecting reading. Levin, Baum, and Bostwick (1965) and Levin and Watson (1965) used lists of PAs in which words composed of novel symbols served as stimuli and familiar English words as responses to study the effects of different letter-sound associations on reading. Bishop (1964) simulated the process of learning to read by teaching college students to read some Arabic terms composed of novel graphic stimuli. These stimuli presumably were as novel to college students as the letters of the alphabet are to beginning readers.

The learner's method of responding and the instructor's method of confirming results are two variables operating both in the beginning stages of reading and in PA learning. In beginning reading, the child can respond by writing or speaking his answer to the printed letter or word. The teacher may provide either visual confirmation (showing him the correct word) or auditory confirmation (telling him). These two response modes and these confirmation modes both have been compared in laboratory studies of PA learning. Conflicting results have been reported. Several authors (Cummings & Goldstein, 1964; McGeoch & Irion, 1952; Otto, 1961) have stated that

it is impossible to establish one method of presenting information as unequivocally superior to another. It is important to note, however, that previous studies of response or confirmation modes have investigated only one or another of these variables. No study has made use of an experimental design that would allow for simultaneous investigation of both response and confirmation modes. As a result, the possibility of an interaction between these two variables has never been explored. Such an interaction conceivably could be used to establish the conditions under which one method of presenting information is superior to another.

The purpose of this study was to investigate the effects of two variables on the learning of a PA task that was similar in many respects to beginning reading. These two variables were (*a*) *S*s method of responding, and (*b*) *E*'s method of confirming results. To investigate these variables, college students were taught to associate a series of novel-symbol–nonsense-syllable pairs in the usual PA paradigm. The symbols used were unfamiliar but discriminable to *S*s. They were presented visually so as to simulate the earlier stages of reading described by Gibson (1965). Responses were either written or spoken; confirmation was visual or auditory. The *S*s' initial learning rate of the PA task and later ability to decode triphoneme "words" were measured.

METHOD

Subjects

The *S*s for this study were 64 male and female undergraduate introductory psychology students at the University of Pittsburgh.

Apparatus[1]

Each *S* was run individually in an experimental room. When *S* was ready, *E* projected the first novel symbol on a Sawyer's Mirascreen 3 feet in front of him, using a Kodak Carousel 800 projector. A Hunter decade interval timer controlled display duration. When the symbol had appeared for 1 second, the projector advanced to a blank slide, and *S* responded, either by writing or speaking his answer. He then pushed the stimulus-control button, bringing either visual or auditory confirmation of results. Visual confirmation consisted of the novel-symbol–nonsense-syllable pair projected on the screen for 1 second by the Carousel projector. Auditory confirmation consisted of simultaneous presentation of the novel symbol on the screen and the sound of the nonsense syllable played back for 1 second by a Bell and Howell Language Master. Visual confirmation duration was controlled by the Hunter timer; auditory display time was controlled by the Language Master. A series of 10 symbol–nonsense-syllable pairs were presented in this fashion and repeated until a criterion of one successful trial was reached.

Procedure

The *S*s were randomly assigned to learn the list of 10 PAs under one of the four following conditions:

1. Written response-visual confirmation. (These *S*s wrote their answers and then viewed the correct response.)
2. Written response-auditory confirmation. (These *S*s also wrote their answers, but heard the correct response.)
3. Spoken response-visual confirmation. (The *S*s spoke their answers, then viewed the correct response.)
4. Spoken response-auditory confirmation. (They spoke their answers and then heard the correct response.)

All *S*s performed two tasks. Task 1: *S*s learned a list of 10 novel-symbol–nonsense-syllable pairs in the usual PA paradigm. During the first trial, each *S* responded when the symbol appeared by pushing the stimulus-control button. This brought either visual or auditory presentation of the correct nonsense syllable. Thereafter, each *S* was taught to respond with the correct nonsense syllable to each of the novel stimuli according to the method prescribed by his experimental condition. Written-response groups wrote their answers on paper, using a separate page for each answer. Spoken-response groups spoke their answers as *E* recorded. The first session ended when *S* completed one errorless trial. Task 2: 10 minutes after the first task was completed, *S*s decoded 10 triphoneme "words," each composed of three symbols learned in the first session. Each *S* used the same response mode as in Session 1.

Analyses of variance were performed for the four groups using response and confirmation modes

[1] The apparatus used in this research was designed and constructed by Robert H. Gibson, University of Pittsburgh. The projector, interval timer, and Language Master were used together in such a manner that stimuli could be presented either by visual or auditory means, or by both means simultaneously. Confirmation could be given by the same three methods. This equipment is useful for studying the effects of different types of stimulus presentation and feedback on learning. It is described in detail in an unpublished study by R. H. Gibson and J. T. Gibson, "A Device to Provide Both Auditory and Visual Feedback in Verbal Learning Studies."

as the independent variables. Dependent variables were trials to criterion in Task 1 and number of errors in Task 2.

RESULTS

An analysis of variance showed that the choice of response and confirmation modes had no effect on trials to criterion in Task 1. Differences did not reach the .05 level of significance. As an additional test for differences between groups in initial learning, the number of overlearning trials (i.e., the number of correct trials for each syllable beyond two correct anticipations) was compared for the four experimental groups. Again, analysis showed no differences. There was no apparent difference whether *S*s spoke or wrote their answers, or whether they saw or heard the correct answer.

When the number of errors made on Task 2 was studied, however, differences were found between the four groups. Analysis of variance showed significant differences in error rate due to the response modes used ($F = 15.65$, $p < .01$). Figure 1 shows that written-response groups made fewer errors than did groups using spoken responses. Analysis of variance also showed a significant interaction between response and confirmation modes in this task ($F = 11.06$, $p < .01$). When *S*s wrote their answers, visual confirmation produced fewer errors than did auditory confirmation. When *S*s spoke their answers, the reverse was true; auditory confirmation produced fewer errors.

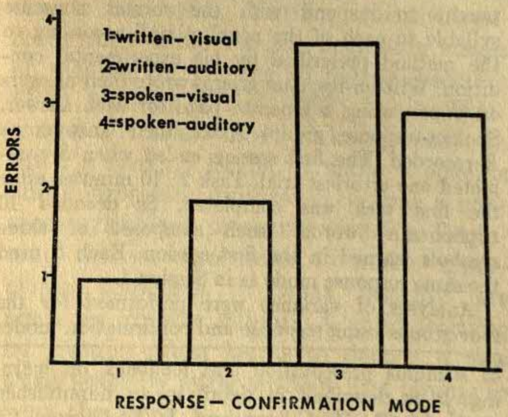


FIG. 1. Mean number of errors in Task 2.

DISCUSSION

It is interesting to note that while apparently all groups required the same number of trials to criterion in Task 1, they responded differently from one another 10 minutes later in Task 2. Two differences in the natures of these tasks may have caused this finding. First, learning rate was measured in the earlier Task 1, while retention was measured later in Task 2. This difference in dependent variables alone may have been responsible for the results. A second explanation lies in the relative complexities of the two tasks. Task 1 required simple sound-symbol associations. Task 2 required decoding or reading of triphoneme "words" composed of the learned symbols. These "words" approximated what Gibson (1965) considered to be higher order units of structure. Most researchers agree that reading higher order units entails more than simple sequential letter-sound associations. It is possible that success in Task 2 required a form of learning that was not measured by the dependent variables in Task 1.

The *S*s who wrote their answers in Task 2 made fewer errors than did *S*s who spoke them. There are two explanations possible for the superiority of the written response in this situation. McGeoch and Irion (1952) suggested that the effect of a particular sense modality may depend on *S*'s familiarity with it. It appears valid to extend this theory to cover familiarity with a particular response mode. In other words, the superiority of the written response may be due simply to the fact that college students have had much more experience during their school careers in writing answers than in speaking them. However, another and different explanation of the superiority of the written response may be made. Monroe (1933) wrote that kinesthetic responses such as "writing in the air" helped remedial readers to discriminate between words. This type of response may have been useful because it focused attention on the accurate spelling of the words and made discrimination easier. In the present study, the written response may have functioned in a similar fashion to make dis-

crimination easier. (Although they recorded no data for this occurrence, *E*s noted that *S*s who spoke their answers often slurred or omitted the final consonant of the nonsense syllables. The *S*s who wrote their answers, on the other hand, were forced to make true letters, and thus perhaps made more careful discriminations.)

Finally, a discussion of the interaction between response and confirmation modes in Task 2 is in order. The *S*s who wrote their answers made fewer errors when visual confirmation was used. When *S*s spoke their answers, auditory confirmation was more effective. McGeoch and Irion (1952) felt that the effect of a particular sense modality was due to the *S*'s familiarity with that mode. It should follow that an *S* given experience at receiving feedback through a particular sense modality would be able to respond more effectively through that same sensory mode. This explanation would serve also to explain why the college students in the written-response–visual-confirmation group made fewer errors than any other group. As regards the implications for teaching reading, the interaction found between response and confirmation modes suggests that the teacher should decide what he wants the student to do before he selects a confirmation mode. Visual means of providing correct answers probably are most effective in teaching reading when achievement is based on writing proficiency. Auditory confirmation probably is more effective when speaking proficiency is desired.

Several major questions still need to be answered. First, were the results obtained in this study due to differential familiarity with the particular systems? In this experiment, *S*s were of college age and were undoubtedly more familiar with the visual-confirmation–written-response system than small children first learning to read. Second, most reading experts agree that reading is more than the association of sounds with words; it involves, in addition, the association of meaning with these sounds and words. A question still remains, therefore, of whether the same results would be obtained with meaningful words rather than the meaningless terms used in this study. A comparison of the response-confirmation modes in the decoding of meaningful terms by adults and young children would answer both these questions. This research currently is being planned.

REFERENCES

BISHOP, C. Transfer effects of word and letter training in reading. *Journal of Verbal Learning and Verbal Behavior*, 1964, **3**, 215–221.

CUMMINGS, A., & GOLDSTEIN, L. The effects of overt and covert responding in two kinds of learning tasks. In J. DeCecco (Ed.), *Educational technology*. New York: Holt, Rinehart & Winston, 1964.

FRIES, C. *Linguistics and reading*. New York: Holt, Rinehart & Winston, 1963.

GIBSON, E. J. Learning to read. *Science*, 1965, **148**, 1066–1072.

LEVIN, H., BAUM, E., & BOSTWICK, S. Cited by E. J. Gibson in Learning to read. *Science*, 1965, **148**, 1066–1072.

LEVIN, H., & WATSON, J. Cited by E. J. Gibson in Learning to read. *Science*, 1965, **148**, 1066–1072.

LEVIN, H., WATSON, J., & FELDMAN, M. Writing as pretraining for association learning. *Journal of Educational Psychology*, 1964, **55**, 181–184.

MCGEOCH, J., & IRION, M. *The psychology of human learning*. New York: Longmans, Green, 1952.

MONROE, M. *Children who cannot read*. Chicago: University of Chicago Press, 1933.

OTTO, W. The acquisition and retention of paired associates by good, average, and poor readers. *Journal of Educational Psychology*, 1961, **52**, 241–248.

PIEKARZ, J. Common sense about phonics. *Reading Teacher*, November 1963, 18–23.

(Received December 4, 1967)

NEW IN 1969

DEVELOPMENTAL PSYCHOLOGY

Editor: Boyd R. McCandless

Developmental Psychology will contain articles which represent the broad range of growth and development and their major associated variables. Chronological age as well as sex, socioeconomic status and effects of physical growth variables (including time of maturation and body build) are all considered relevant developmental variables.

Although the adolescent and the aging population appear to deserve particular attention at this time, *Developmental Psychology* will indeed cover the entire span. Cross-species articles and articles concerning developmental research with retardates will also be included.

The articles in *Developmental Psychology* will not be limited to any one approach such as "experiments versus studies" or "cognition versus motivation".

Bimonthly. One volume per year.
Subscription Price: $10.00 per year

AMERICAN PSYCHOLOGICAL ASSOCIATION
1200 Seventeenth Street, N. W.
Washington, D.C. 20036

HOLT, RINEHART AND WINSTON, INC.

CRITICAL MOMENTS IN TEACHING... Five New Films

Critical Moments in Teaching is the exciting new series of 16 mm. color films that fills an urgent need in teacher education. It provides the teacher-in-training with vital experience in making decisions about actual teaching problems. Each film describes a problem which might confront either a secondary or elementary school teacher—such as the basis on which students should be graded or the motivating of a class. The problem itself is elaborated in detail, but no solution is given or implied.

The series can be used as a stimulus for teacher-directed class discussion, role-playing, independent study of analysis, or professional analysis.

Walls

the problem of teaching an unresponsive, passive, high school class. color, 10½ min., $125.00.

A Child who Cheats

handling a problem of cheating in the classroom. color, 10 min., $115.00.

The First and Fundamental R

the problem of teaching reading to elementary school children in a low socio-economic area. color, 12 min., $145.00.

I Walk Away in the Rain

the problem of motivating a highly capable adolescent who puts forth minimal effort in school work. color, 11 min., $130.00.

Image in a Mirror

working with an elementary school child who is lacking in self-confidence. color, 8½ min., $100.00.

For more information write Paul B. Foster, Media Department

ALSO FROM HRW

THE COMPLEXITIES OF AN URBAN CLASSROOM
An Analysis toward a General Theory of Teaching

By Louis M. Smith and William Geoffrey, both of Washington University
1968 — 288 pp. — $5.95

SCHOOL LEARNING
An Introduction to Educational Psychology

By David P. Ausubel, The City University of New York, and Floyd G. Robinson, The Ontario Institute for Studies in Education
April 1969 — 576 pp. — $8.50 (tent.)
Readings — 320 pp. — $4.95 paper (tent.)
Instructor's Manual
Student Workbook

HOLT, RINEHART AND WINSTON, INC.
383 Madison Avenue, New York 10017

Announcing

ADULT PSYCHOLOGY

Ledford J. Bischof

The first book to deal exclusively and comprehensively with the universal theories and research on the average (not deviant) adult, from early adulthood to old age. Intended for undergraduate or graduate courses, it is ideal as a complement to developmental texts. Unique features: An entirely new approach to maturity and self-image and a wealth of data on the physiological processes in adulthood. Ready in March. *$5.00 (tentative) paper*

Other Texts for Your Consideration

STIMULUS AND RESPONSE

John A. Barlow

In programmed form, a development of contiguity theory and its application to some fundamental elements of operant psychology. Instructor's Manual. *199 pp.; $4.75 paper*

PSYCHOLOGICAL FOUNDATIONS OF EDUCATION
SECOND EDITION

Morris L. Bigge and Maurice P. Hunt

A well-documented, problem-oriented text taking a semihistorical and comparative approach to the basic topics of how children develop through adolescence, how they learn, the relationship between development and learning, and how knowledge of this relationship promotes more effective teaching. Instructor's Manual. *603 pp.; $9.95*

PSYCHOLOGY IN THE CLASSROOM, SECOND EDITION

Rudolf Dreikurs

This practical manual, grounded in the philosophy of democracy and the socio-teleological approach of Adlerian psychology, provides background information and methods necessary to deal with behavior problems and learning deficiencies of students. *286 pp.; $3.75 paper*

REPORT WRITING IN PSYCHOLOGY AND PSYCHIATRY

Jack T. Huber

Covers approaches to formulating a case, outlines for formulation, outlines for reports in special areas, confidentiality, therapy progress notes, techniques of writing, supervision, and examples of report writing. *114 pp.; $2.50 paper*

LEARNING AND HUMAN ABILITIES: EDUCATIONAL PSYCHOLOGY, SECOND EDITION

Herbert J. Klausmeier and William Goodwin

Emphasizes the concept of emerging human abilities, thus integrating the treatment of growth and learning. Student Evaluation Guide. *720 pp.; $9.50*. Student Workbook—*$3.25*

1817

Harper & Row, Publishers · 49 East 33d St., N. Y. 10016

TITLES IN PSYCHOLOGY FROM MACMILLAN

General Psychology

By David C. Edwards, Iowa State University

This is a brief yet thorough introduction to psychology—its content, its terminology, and its methods. The author has accomplished brevity by carefully selecting the important topics, and by avoiding unnecessary or irrelevant exposition and discussion of the obvious. The approach is modern, and is concerned with psychology as a science of behavior. The text is suitable for either a brief or a standard course in general psychology in which supplementary materials are used. A Teacher's Manual is available gratis.

1968, 394 pages, $5.95

Workbook to accompany *General Psychology*

1968, 115 pages, paper, $1.95

Educational Psychology
Third Edition

By Glenn Myers Blair, R. Stewart Jones, and Ray H. Simpson, all of the University of Illinois

This successful textbook, now in its third edition, gives students and graduate teachers practical, up-to-date help in managing the learning process in keeping with sound psychological theory. It contains much new material—especially in the sections on teacher self-appraisal, the psychology of adolescence, the social psychology of teaching, and diagnostic and remedial procedures. Employing a developmental emphasis throughout, the authors examine the growth of the child; the use of psychological tools to evaluate educational programs; and some of the factors which influence the professional growth and mental health of the teacher. A Teacher's Manual is available, gratis.

1968, 704 pages, $8.95

Readings in Educational Psychology
Second Edition

Edited by Victor H. Noll and Rachel P. Noll, both, Michigan State University

This thoroughly revised collection of readings provides the student in educational psychology with a source of significant and relevant literature in the field. Twenty-eight of the articles are new and reflect a changing emphasis in educational psychology. Especially written for this compilation are articles by Robert C. Craig, Robert L. Ebel, Ruth Strang, Elizabeth M. Drews, and William Gnagey.

1968, 432 pages, paper, $4.50

An Introduction to Educational Research
Third Edition

By Robert M. W. Travers, Western Michigan University

Much material in the third edition of this fine text has been rewritten to achieve a simpler presentation throughout. The entire book has been rendered increasingly useful to the Master's Degree candidate by the elimination of much highly technical material and by the inclusion of new chapters on computers, content analysis, and classroom devices for recording and analysis. The theme remains, however, that only educational research based on theory will have lasting results. The text is organized around the main approaches to research in the behavioral sciences: survey methods, prediction methods, experimental methods, historical methods, case studies and environmental methods.

1969, approx. 512 pages, prob. $7.95

Write to the Faculty Service Desk for examination copies.

THE MACMILLAN COMPANY **866 Third Avenue** **New York, New York 10022**

In Canada, write to Collier-Macmillan Canada, Ltd.,
1125B Leslie Street, Don Mills, Ontario

Scott Foresman College Division

Applying Theory To Practice

PSYCHOLOGY AND TEACHING

Third Edition

By Norman T. Bell, Michigan State University
David D. Starks, University of Michigan

This thoroughly revised study is a basic text for an educational psychology course, a useful guide for teachers, and an excellent supplementary reference for courses in teaching. The overall organization consists of discussion on the role of the teacher, the nature of children, and the nature of learning. The general approach of the text stems from the application of psychological data to the operation of a secondary or elementary classroom and the combination of psychological principles with the consideration of actual teaching situations. The plan is emphasized by directing special attention to the group aspects of educational psychology, classroom discipline procedures, and the problems of disadvantaged children.

Extensive measures are taken to facilitate the assimilation of material. A brief introduction to each chapter identifies the major conceptual materials and important principles presented in the discussion. Opportunity to apply learning to actual classroom situations occurs periodically in the reading through the addition of updated case studies. Special programmed chapters are included which contain charts to be filled in by the student. These charts culminate at the conclusion of the work in a comprehensive summary of all the material. In addition, self-testing exercises appear at the end of each programmed unit and frequent reference to preceding material, integrated into the flow of the text, allows the student to review, test, and reinforce his learning. All materials have been tested on samples from the student population and edited on the basis of that testing program.

This edition contains twelve completely new readings, a rewritten reference manual on general psychology, use of new references and suggestions for further reading. Psychology and Teaching avoids emphasizing any one cause in favor of considering all aspects of educational psychology.

Ready Spring 1969, approx. 768 pages, prob. $9.75.

SCOTT, FORESMAN AND COMPANY College Division

Glenview, Illinois Atlanta Dallas Palo Alto Oakland, N.J.

JOURNAL OF
EDUCATIONAL PSYCHOLOGY
MONOGRAPH SUPPLEMENT

Vol. 59, No. 6, Part 2 December, 1968

PRODUCTION AND JUDGMENT OF SOLUTIONS TO FIVE PROBLEMS[1]

DONALD M. JOHNSON, GEORGE L. PARROTT, AND R. PAUL STRATTON
Michigan State University

Different groups of college *Ss* wrote one solution or many solutions to problems, like the plot-title problem, of verbal, numerical, and pictorial material. Instructions to write many solutions yielded solutions of lower mean quality but more superior solutions. Information about criteria for good solutions raised quality. Large quantity was associated with low quality, both for variations in conditions and for individual differences within conditions. Differences between problems in the quality-quantity relation were dependent on the number of superior solutions to the problem. 3 types of judgment training—individual, dyadic, and tutorial—interpolated between production of solutions and selection of the best solution were generally successful and, under certain favorable conditions, improved overall performance.

Problem solving, like learning, is not a simple homogeneous activity. This truism is important for the analysis of problem solving because a statement that holds for one component process may not hold for another. It is important also for attempts to improve problem solving because a procedure that facilitates one process may not facilitate another. Therefore, the research to be described separates problem solving into three different but functionally interdependent processes: preparation, production, and judgment.

Intellectual tasks begin with some kind of preparation, most often the acquisition and organization of information, as by listening to instructions or by reading a printed paragraph. After preparation, some tasks are primarily productive, as in writing many uses of a brick; and some are mostly matters of judgment, as in selecting the best answer on a multiple-choice test. The present research is concerned with that large class of problems, not often investigated, for which both production and judgment are required. The *S* produces several possible solutions, then examines these and picks one as his best effort. In such cases production is different from preparation but depends on it. Judgment is different from production, but it is the solutions produced that are judged. The different processes are interdependent, but they can be studied separately, and conditions that influence each can be experimentally manipulated.

Early attempts at analysis of thinking into component processes (Dewey, 1910; Wallas, 1926) suffered from the inadequacies of a subjective method. More objective analyses of protocols obtained from poets and artists by the thinking-aloud method have been published (Patrick, 1935, 1937). Similar procedures have been used to study the problem-solving processes of students at work (Bloom & Broder, 1950; Burack, 1950) and to collect data for computer simulation of thinking (Newell, Shaw, & Simon, 1958). Another procedure (Johnson, 1960, 1961) maintains more control

[1] This investigation was supported by the U. S. Office of Education, Project No. 5-0705. A final report containing several supplementary studies will be available in mimeographed form.

Copyright © 1968 by the American Psychological Association, Inc.

over S's activities by serial exposure of the materials of the problem. This has permitted identification and timing of two or three operations, and study of the contribution of each to the whole problem-solving enterprise with only slight disturbance of overall performance. Factor analysis of individual differences in success on minor problems often leads to factors that are labeled as problem-solving processes (Guilford, 1967), but the technique can also be applied to time spent on each process (Johnson & Jennings, 1963). Records of time spent on preparation, production, and judgment by college students doing three plot-title problems yielded low correlations between different processes and high correlations between like processes. Aside from the specific findings these studies have demonstrated the feasibility of analysis of the solution of problems into a few processes by manipulation of instructions and exposures and the study of their interrelations.

Brainstorming, although aimed at practical rather than analytical goals, depends on an implicit analytical assumption. The hypothesis that deferment of judgment improves production of ideas rests on the assumption that judgment and production are different processes that can occur at different times. (The other hypothesis of brainstorming, that social interaction facilitates production, is not relevant to the present research.) Instructions and training are directed toward freeing S from premature self-criticism. Meadow and Parnes (1959) found that a 30-hour course in creative problem solving did increase quantity and quality of output on certain tasks such as Guilford's test of producing unusual uses for common objects. Meadow, Parnes, and Reese (1959) compared brainstorming instructions, emphasizing quantity of production and ignoring quality, with nonbrainstorming instructions, emphasizing quality and penalizing for ideas rated poor. More solutions of high quality came from the brainstorming group. Parnes and Meadow (1959) also found a difference in favor of brainstorming instructions and, in addition, a high correlation between quantity and quality. That is, those Ss who produced more solutions produced more superior solutions. Later, Parnes and Meadow (1960) tested some of their students 8 months after a course in creative problem solving and found improvement on all tests, including plot titles rated for high quality, as compared with control groups.

A rather detailed study of instructions by Gerlach, Schutz, Baker, and Mazer (1964) raised questions that must be considered in further research. They wrote six sets of instructions, including brainstorming, nonbrainstorming, penalty for bad responses, and what they called "criteria-cued" instructions. The instructions for this group read: "The more imaginative or creative your ideas, the higher your score will be. Each idea will be scored in terms of how unique it is, how valuable it is . . . the more original and creative the better." On the familiar test of writing uses for a coat hanger the criteria-cued instructions did yield the most good responses, hence the authors argue that the improvement attributed to brainstorming could be due to learning a criterion of quality. This interpretation, which is quite different from the original conception of brainstorming, suggests that the improvement may be in the judgment process as well as the production process.

Records of production over time should help to explain total output. It is conceivable that S produces his best solution on the first try and, if pushed to produce more, produces solutions of lower quality. It is also conceivable that the good solutions appear later. To date, the results have been conflicting. Christensen, Guilford, and Wilson (1957) found that the production of simple responses decreases during a working period of about 15 minutes while the production of plot titles was linear and the quality of these titles was constant. Johnson and Jennings (1963) had their Ss write five plot titles and found that the best one occurred in each position equally often. Parnes (1961) found an increase in the number of unusual uses and takes this result as an argument for extended effort in brainstorming. Apparently the course of

production is different for simple responses and for complex solutions, but further research is needed.

This brief review of relevant experiments points up the need for a fundamental examination of the interrelated contributions of production and judgment to problem solving. The results obtained to date suggest certain methods to be tried and certain errors to be avoided. An adequate description of the processes leading to the final solution requires detailed analysis of a large number of solutions and some additional variations in conditions. Curiously, although the outputs of *S*s producing many solutions under different instructions have been compared, the standard condition has been overlooked. Instructing *S* to write one solution to a problem may be considered the standard condition to which more complicated conditions should be compared. The question of how well *S*s can judge their own solutions must be examined because assumptions about self-criticism are commonly included in speculations about thinking. A previous study (Johnson & Jennings, 1963) with limited data indicated that the accuracy of college students in judging their own solutions to the plot-title problem was above chance levels but not very high. The possibility of increasing accuracy in judgment is worthy of serious investigation since improvement has been obtained under some conditions (Johnson & Zerbolio, 1964).

The measures of interest are the number of solutions produced by each *S*, the average quality of these solutions, and the number of superior solutions. Differences between *S*s on these measures may also be enlightening. Reliable ratings of the solutions are necessary; other methodological considerations, suggested by reports of previous research, will be mentioned below.

The first experiments analyzed the contributions of production and judgment to problem solving. Since the results emphasized the role of judgment, brief programs for training judgment were prepared, and the later experiments evaluated the improvement obtained. The same problems and some of the same procedures were employed in all experiments.

Standardization of Problems

The present research is focused on problems with many solutions that cannot be dichotomized as right or wrong but can be graded in respect to such qualities as usefulness, appropriateness, cleverness, and originality. This might be called productive thinking as well as problem solving. In Guilford's (1967) terminology it is divergent thinking as opposed to convergent thinking. Another consideration was that the problems should be substantial problems that would offer some challenge to college students. Tasks as simple as writing uses for a brick or giving uncommon associations to words have been criticized as trivial. Any punster knows that the set for simple verbal productions is highly vulnerable to variations in instructions and social atmosphere.

It was necessary to choose problems yielding solutions that vary considerably in quality and that can be reliably rated as to quality. Finally, within the restrictions of the design, the problems should vary in content. The literature on problem solving contains many findings that apply, as far as is known, to only one problem. The use of several problems permits general principles to emerge as well as differences between problems.

The plot-title problem, which has been used in some of the research mentioned above, meets these specifications; and it is also a good test of the originality factor, according to Wilson, Guilford, and Christensen (1953). Reading the instructions for the problem and then the plot itself may be considered the preparation. Production consists of writing titles for it. Judgment consists of selecting the best of these as the final solution.

Twelve problems were constructed along the lines of the plot-title problem, two examples each of six types, and tried out with several samples of college students. Four were chosen on the basis of the number of solutions written in 5 minutes, the correlation between ratings of the solu-

tions by two raters, and the standard deviation of the ratings. Two other types, Consequences and Plot Completions, did not yield satisfactory interrater agreement. The problems chosen, as well as the plot-title problem, are briefly described below, together with instructions for multiple solutions.

Plot Titles. A paragraph gave the plot of a story or movie, with the following instructions:

"Your task is to think of titles for the story. Read the plot then write as many titles for it as you can."

Table Titles. A table of agricultural data, showing four columns of statistics for seven time periods, was printed, with the following instructions:

"Table X below has reference to United States statistics, and was taken from a past volume of the *World Almanac.* You are to examine the table and then write as many titles for it as you can."

Conclusions. A chart was printed, with column diagrams representing social-welfare expenditures under five public programs for six time periods, with the following instructions:

"Figure XVI is taken from *The Statistical Abstracts of the United States, 1964.* What can you conclude from this table? Write short sentences, as many as you can, each of which summarizes a generalization from this table."

Sentences. Wallach and Kogan (1965) asked children to write short stories using four words. For present purposes with adult *S*s the integration of four words in a single sentence seemed preferable:

"Write many sentences, each of which contains these four words."

happy expensive horse lake

Cartoon. A cartoon was presented in four squares, with the printing removed from the last square. The instructions were to "write as many different quotes for the last square as you can."

Although the solutions were all written in words, the materials presented included verbal, numerical, and pictorial materials in order to achieve some variety of content.

EXPERIMENT I: PRODUCTION AND JUDGMENT

Method

Instructions

When *S*s are instructed to ignore quality or not to be critical, they write more solutions, but the qualitative and quantitative effects of the instructions are confounded. Hence Group 1 was instructed to write one solution to each problem, and Group 2 was instructed to write as many as possible. There was no reference to quality in either case.

Group 3 was instructed to write as many solutions as possible, then to select the best. This variation was planned as a check on *S*s' ability to judge their own solutions and as a comparison between preferred solutions and the single solutions produced by Group 1.

Groups 4 and 5 were included to examine the effects of criteria-cued instructions reported by Gerlach et al. (1964). The instructions asked for "good" or "clever" solutions, and these were followed by more specific criteria based on the criteria developed by the raters:

Plot Titles. By clever we mean an imaginative, creative, or unusual title for this plot.

Table Titles. A good title is a comprehensive one that includes the important points concisely.

Conclusions. A good conclusion would be a valid generalization which integrates the table as a whole.

Sentences. A good sentence reads smoothly; the four words fit unobtrusively into the structure of the sentence.

Cartoon. A clever quote is an imaginative idea that fits the cartoon.

Group 4 received the criteria-cued instructions, along with a request to write as many as possible. Group 5 received the same instructions, and, like Group 3, was requested to select the best solution later.

Procedure

Since each *S* was to do five problems, the possibility of order effects arose. In Groups 3 and 5, especially, after *S* did the first problem and discovered that postproduction judgment was required, performance on later problems could be influenced. Therefore five orders were arranged, with each problem appearing once in each position.

Thus 25 types of booklets were prepared: separate forms for the five groups, and five orders for each form. Each problem was printed on a page of 8½ × 11 inch paper; the different orders were arranged when the pages were stapled in booklets. Following each of the five problem pages, Groups 3 and 5 had additional pages with the instructions: "Now turn back to the titles (or sentences, etc.) you wrote and pick the best one. Put a check mark (✓) beside it." Other groups had filler pages of irrelevant material expected to require the same working time.

Since some of the standardization *S*s complained that 5 minutes did not suffice, 7 minutes were allowed for each problem.

Subjects

The *S*s were 200 students in general psychology at Michigan State University, mostly freshmen and sophomores, divided into five groups of 40 each, within which there were five orders. Eight each of the 25 types of booklets were distributed serially during a regular class meeting. Eight booklets were returned incomplete, so eight *S*s, drawn from the same population, were run later to fill the missing cells.

Scores on the College Qualification Test (CQT) and a locally constructed reading test were available for most students. Table 1 shows that the five groups were quite similar in respect to these scores. (Since the scores for two *S*s could not be found, the means of Groups 1 and 3 are based on only 39 scores).

Results

The results consist of 5,215 solutions, about 1,000 for each problem. Each solution was typed on one side of a card, and code numbers were typed on the other side. The cards for each problem were shuffled and given to two judges for blind rating.

TABLE 1

MEANS AND STANDARD DEVIATIONS OF FIVE GROUPS ON TWO TESTS

Group	Reading Test		College Qualification Test	
	M	*SD*	*M*	*SD*
1	29.9	6.7	128.5	26.3
2	30.6	5.8	131.8	18.6
3	31.7	6.4	137.4	22.8
4	31.9	8.0	133.7	27.5
5	30.3	6.2	129.3	21.1

Agreement Between Judges

The judges had rated the solutions obtained earlier and had cooperatively written general criteria and specific points on a scale of 1 (low quality) to 7 (high quality) for each problem. They used solutions from incomplete papers for additional practice, discussion, and refinement of ratings. After this practice, amounting to about 30 solutions per problem, they rated all solutions independently, one problem at a time. The two ratings, together with code numbers, were then punched on cards for electronic data processing.

Since the reliability of the ratings is crucial for this type of research, the solutions were rated in blocks of about 200 and interjudge correlations were computed for each block. Table 2 shows that Conclusions and Sentences were the easiest to rate, but in general the agreement was adequate and fluctuations over time were small. (The last correlation in each line of Table 2 is based on less than 200 solutions.) Subsequent computations use the sums of these ratings, which range from 2 to 14.

TABLE 2

CORRELATION COEFFICIENTS COMPUTED BY BLOCKS OF 200 SOLUTIONS TO SHOW INTERJUDGE AGREEMENT THROUGHOUT RATING PERIOD

Problem	Block						
	1	2	3	4	5	6	7
Plot Titles	.548	.688	.596	.746	.764	.747	.673
Table Titles	.988	.988	.986	.984	.974		
Conclusions	.748	.908	.922	.924	.922		
Sentences	.866	.808	.822	.808	.787		
Cartoon	.658	.748	.748	.795	.740	.787	

Order Effects

The data for each problem were inspected for order effects, but no such effect appeared, either for number of solutions or mean rating of the solutions. Hence the data obtained from the five positions have been combined in later analyses.

Number of Solutions Produced

Each *S* of Group 1 wrote one solution as instructed, but the frequencies in the four multiple-solution groups, shown in Table 3, were influenced by the more specific in-

TABLE 3

MEAN NUMBER OF SOLUTIONS PER SUBJECT TO FIVE PROBLEMS PRODUCED BY FIVE GROUPS

Group	Plot Titles	Table Titles	Conclusions	Sentences	Cartoon	Total
1	1.00	1.00	1.00	1.00	1.00	1.00
2	9.42	5.58	6.58	6.00	6.75	6.86
3	8.72	5.68	6.78	6.45	7.40	7.00
4	6.57	4.88	5.05	5.10	6.35	5.59
5	7.80	4.92	4.55	4.72	6.08	5.62
Total	6.70	4.41	4.79	4.66	5.52	5.22

TABLE 4

FREQUENCIES OF RATINGS, ON SCALE OF 2–14, OF SOLUTIONS TO FIVE PROBLEMS

Problem	Rating												
	2	3	4	5	6	7	8	9	10	11	12	13	14
Plot Titles	140	125	249	217	232	146	120	52	30	18	4	4	4
Table Titles	145	15	93	10	73	15	209	13	117	5	124	1	62
Conclusions	123	25	59	36	247	54	75	55	123	46	68	24	23
Sentences	39		34	38	256	151	205	81	83	26	9	5	4
Cartoon	55	31	93	78	181	166	239	137	79	29	20	8	1
Total	502	196	528	379	989	532	848	338	432	124	225	42	94

structions. In contrast to Groups 2 and 3, Groups 4 and 5 received the criteria for good solutions. In contrast to Groups 2 and 4, Groups 3 and 5 had to identify their best solutions. The effects of these variations on each problem were tested by a 2×2 analysis of variance, with 40 *S*s in each cell. The criteria-cued instructions significantly reduced productivity on Plot Titles ($p < .01$), Conclusions ($p < .01$), Sentences ($p < .01$), and Cartoon ($p < .025$). The reduction was not significant on Table Titles, but it was significant when each *S*'s total production on all five problems was treated as a single score ($p < .01$). Instructions to select the best solution did not influence productivity.

Distribution of Ratings

Table 4 shows the distributions of the ratings of the solutions to the different problems for all five groups combined. Some positive skew is apparent. The judges attempted to spread out their ratings but, as one might expect, the ratings piled up at the low end.

The irregularity that appears in the distributions should be noted. When interjudge agreement is high, the sum of two ratings is usually an even number. Odd sums occur only when raters disagree. The greater frequency of even sums over odd sums appears most sharply in solutions to Table Titles and Conclusions because the reliability of the ratings was very high in each case. This effect has no bearing on the present research.

The consequences of the production of many solutions can be best understood by examination of complete distributions of ratings. The clearest comparison is between Groups 1 and 2, which differed only in number of solutions produced, hence Figure 1 has been prepared to facilitate this comparison for each of the five problems. The irregularity mentioned above has been removed by combining even and odd sums, with the exception that the highest ratings, 12, 13, and 14, have been combined in one interval. In general, when *S*s were instructed to write many solutions, they wrote more superior solutions, more mediocre solutions, and more inferior solutions. Since they wrote more solutions of all degrees of quality, they wrote more solutions of high quality.

Intraindividual Variability

A fundamental characteristic of productive thinking under instructions to produce many solutions is the variation in quality of the solutions produced by each *S*. It is conceivable that some *S*s might write mostly inferior solutions, others mostly mediocre solutions, and others mostly superior solutions, but the raw data do not show such results. For example, the printout of the ratings of each of the eight *S*s of Group 2 who had the plot-title problem in the first position show that the individual ratings extended, respectively, from 2 to 8, 2 to 7, 2 to 11, 2 to 11, 2 to 13, 4 to 7, 5 to 9, and 6 to 7. A tabulation of the ranges for the 160 *S*s of Groups 2–5 on each of the five problems, taken one at a time, is displayed in Table 5. Since the complete range of

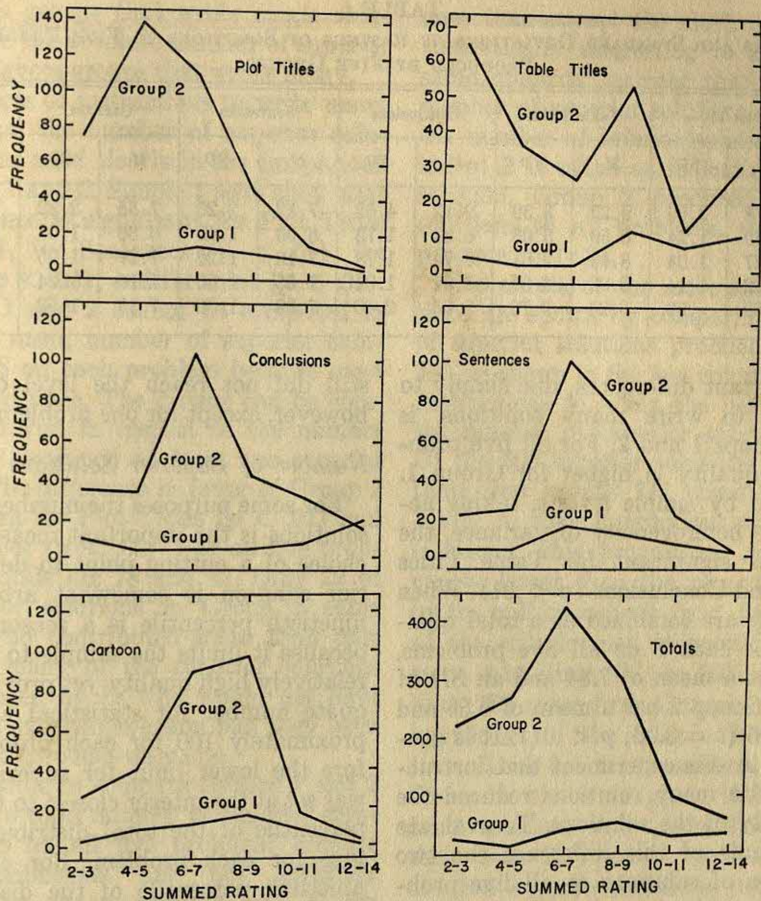


FIG. 1. Distributions of ratings of solutions to five problems by 40 *S*s who wrote one solution (Group 1) and 40 *S*s who wrote many solutions (Group 2). (The base line represents the sum of two independent ratings on a scale of 1–7. Odd and even sums have been combined to remove irregularities, except that the highest ratings, 12, 13, and 14, have been placed in one interval.)

the ratings was 13 (2–14), Table 5 demonstrates that the variability in quality of the solutions produced by single *S*s was large. Almost all *S*s who wrote more than one solution wrote solutions of a wide range of quality. The median range per *S* per problem was about 7. (Only 786 sets of solutions are represented in Table 5 because 14 *S*s wrote only one solution.)

Mean Quality of Solutions

In order to compare the multiple-solution groups with the single-solution group, each *S* in the multiple-solution groups was assigned a score representing the mean of the ratings of all his solutions to a problem. It is the mean of these 40 means for the 40 *S*s that is entered in each cell in Table 6. The standard deviations of these distributions of mean scores are generally smaller than the standard deviations of the distributions of single ratings for Group 1 because the averaging reduces the variability.

TABLE 5

INTRAINDIVIDUAL RANGES IN RATINGS OF SOLUTIONS

Range	1	2	3	4	5	6	7	8	9	10	11	12	13
Frequency	17	25	77	76	117	91	138	79	59	39	49	6	13

TABLE 6

MEANS AND STANDARD DEVIATIONS OF RATINGS OF SOLUTIONS TO FIVE PROBLEMS PRODUCED BY FIVE GROUPS

Group	Plot Titles		Table Titles		Conclusions		Sentences		Cartoon		Total	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	5.78	2.66	9.22	3.39	9.12	4.62	7.52	2.26	7.55	2.22	7.84	1.80
2	5.49	1.24	6.59	3.03	6.83	1.73	6.99	1.42	6.86	1.01	6.56	.95
3	5.27	1.04	8.13	2.61	6.74	1.78	7.19	1.26	7.27	1.25	6.90	.77
4	5.72	1.14	8.22	3.01	7.91	1.94	7.55	1.55	7.16	1.24	7.31	1.10
5	5.53	.99	8.23	2.78	7.65	2.40	7.99	1.06	7.13	1.52	7.31	.89

The important differences, due simply to instructions to write many solutions, is between Groups 1 and 2. For all five problems mean quality is higher for Group 1. When tested by simple t tests, taking account of the heterogeneity of variance, the difference is significant for Table Titles ($p < .01$) and Conclusions ($p < .01$). When these ratings are combined as a total quality score for each S on all five problems, Group 1 has a mean of 7.84 and an SD of 1.80, while Group 2 has a mean of 6.56 and an SD of .95 ($t = 3.98$; $p < .01$). It is generally true in this experiment that instructions to write many solutions reduced the mean quality of the solutions. To evaluate the magnitude of this difference the two distributions of solutions to all five problems were combined as one distribution of 1,573 solutions. In respect to this combined distribution (with positive skew), the Group 2 mean was at the fifty-ninth percentile while the Group 1 mean was at the seventy-seventh percentile.

The effects of the more specific instructions given to Groups 2–5 were evaluated by a 2 × 2 analysis of variance for each problem with the 40 means for *S*s in each cell. The request to select the best solution had no significant effect on any problem. The criteria-cued instructions, however, increased mean quality on all five problems. This effect was significant at the .01 level for Conclusions and Sentences and, of course, for total solutions. Thus information about the criteria for good solutions not only reduced the number of solutions produced but also increased the average quality of these solutions. Average quality still did not reach the level of Group 1, however, except for one problem, Sentences.

Number of Superior Solutions

For some purposes the number of superior solutions is the important measure, but the choice of a cutting point to define a superior solution is somewhat arbitrary. The ninetieth percentile is a reasonable choice because it limits the sample to solutions of relatively high quality yet provides an adequate number for statistical analysis, approximately 100 for each problem. Therefore the lower limit for superior solutions was set at the integer closest to the ninetieth percentile of the total distribution of ratings for each problem. For example, the ninetieth percentile of the distribution of ratings of plot titles (see Table 4) fell between 8 and 9. Hence solutions rated 9 or above were considered superior. There were 112 superior solutions by this definition, slightly less than the ideal 10%.

Table 7 shows that seven of these superior solutions to the plot-title problem were produced by the 40 *S*s of Group 1, writing single solutions, and 29 by the 40 *S*s of Group 2, writing many solutions. In general, the multiple-solution groups produced more superior solutions than the single-solution group, though Conclusions is a possible exception.

The percentages in Table 7 were calculated with the number of solutions for the group, shown in Table 3, as the base. That is, the 29 superior solutions to the plot-title problem produced by Group 2 were 8% of the 377 solutions produced by that group. This measure, like the mean quality meas-

ure, favors groups that wrote single solutions, while the absolute number of superior solutions favors groups that wrote many.

The effects of instructions to write many solutions on the number of superior solutions appear most clearly in the comparison of Groups 1 and 2. Simple t tests show that the difference is significant for Plot Titles ($p < .01$), for Sentences ($p < .01$), for Cartoon ($p < .05$), and for total ($p < .01$).

Groups 1 and 2 cannot be compared in respect to mean number of superior solutions per *S* on each problem because most *S*s produced none. The groups can be compared, however, in respect to the number of *S*s who produced at least one superior solution. The difference in favor of Group 2 is significant for Plot Titles ($p < .01$) and for Sentences ($p < .05$). For all five problems the totals are 78 and 47. Only 22 of the superior solutions in Group 2 came from *S*s who contributed more than one. Evidently many of the superior solutions were produced by *S*s who would not have produced a superior solution had they been instructed to produce only one.

It is possible to compare Groups 1 and 2 in respect to total number of superior solutions per *S* on all five problems because these distributions are not seriously skewed. The mean number per *S* for Group 1 was 1.17, with an *SD* of .86. The mean for Group 2 was 2.50, with an *SD* of 1.69, and the difference between the two means is significant ($p < .01$). All these measures indicate that *S*s produce more superior solutions when they are told to write many.

Although the superior solutions are the ones that attract the most attention, it is theoretically interesting to note that the same quantity increase that increases the number of superior solutions also increases the number of inferior solutions. If a rating of 2 is taken as indicating an inferior solution, Group 2 produced 171 inferior solutions to all problems, while Group 1 produced only 23.

The effects of the more specific instructions are shown by comparing the number of superior solutions produced by Groups 2–5, as shown in the last column of Table 7. Instructions to select the best do increase the number of superior solutions slightly. Of the 506 superior solutions produced by these four groups 275 came from Groups 3 and 5, with instructions to judge their solutions. This proportion, .55, is significantly larger than the proportion implied by the null hypothesis ($p < .05$). The effect of the criteria-cued instructions is of the same magnitude. Of the 506 superior solutions 280 came from Groups 4 and 5, with knowledge of the criteria for good solutions. Thus the best condition was that of Group 5 who were given the criteria for good solutions and instructions to select their best. This group contributed 149 of the 506 superior solutions, which is 29% rather than 25%. The finding by Gerlach et al. (1964) of the value of criteria cues is thus confirmed both for mean quality and for number of superior solutions.

Supplementary analyses were carried out using a 95% level as defining a superior solution. The numbers were smaller and the significance levels were lower, but the

TABLE 7

Frequencies (f) and Percentages of Superior Solutions by Groups and Problems

Group	Plot Titles		Table Titles		Conclusions		Sentences		Cartoon		Total	
	f	%	f	%	f	%	f	%	f	%	f	%
1	7	18	5	12	20	50	7	18	8	20	47	23
2	29	8	9	4	17	6	24	10	21	8	100	7
3	28	8	12	5	19	7	25	10	42	14	126	9
4	18	7	20	10	32	16	34	17	28	11	132	11
5	30	10	17	9	27	15	37	20	38	16	149	13
Total	112	8	63	7	115	12	127	14	137	12	554	11

same picture of the results emerged. None of the above statements were contradicted.

Accuracy of Judgment

Those who write many solutions write more superior solutions, but they write more inferior solutions as well. The critical question is one of judgment. Will *S* be able to select his best? If he writes superior solutions and selects his worst or even his average solution, nothing is gained. Table 8 displays the pertinent comparisons for Group 3, with instructions to write many and then choose the best, and for Group 5, with the same instructions plus information about the criteria for good solutions. In each comparison of this table there are 39 or 40 preferred solutions for each problem—a few *S*s did not identify their best—while the number of nonpreferred solutions ranges from 141 to 310. Each *S* was given a score representing the difference between the rating of his preferred solution and the mean rating of his nonpreferred solutions, and the *t* tests indicate whether the means of these difference scores are significantly greater than zero. It is clear from Table 8 that the solutions selected as best were better, by and large, than the others and that in some cases the differences were quite small.

Comparison of the number of superior solutions among the preferred and the nonpreferred solutions, as shown in Table 9, is complicated by the quantity effect; there are more superior solutions where there are more solutions.

TABLE 8

Mean Ratings of Preferred and Nonpreferred Solutions in Two Groups

Problem	Group 3			Group 5		
	Preferred	Nonpreferred	t	Preferred	Nonpreferred	t
Plot Titles	5.47	5.21	.97	6.23	5.42	2.80**
Table Titles	9.26	7.49	3.92**	9.41	7.84	3.48**
Conclusions	7.30	6.62	1.41	8.52	7.39	1.43
Sentences	7.19	7.19	.00	8.10	7.56	2.06*
Cartoon	7.87	6.91	2.41*	7.85	7.08	1.77*
Total	7.42	6.68	1.83*	8.02	7.06	2.20*

* $p < .05$.
** $p < .01$.

TABLE 9

Frequency of Superior Ratings among Preferred and Nonpreferred Solutions in Two Groups

Problem	Group 3		Group 5	
	Preferred	Nonpreferred	Preferred	Nonpreferred
Plot Titles	4	24	4	26
Table Titles	4	8	7	10
Conclusions	6	13	13	14
Sentences	5	20	8	29
Cartoon	10	32	10	28
Total	29	97	42	107

It should be noted that *S*s of Group 5 had information about criteria when they produced their solutions. The difference between Group 5 and Group 3 is due to the influence on production of information about the criteria. The difference between preferred and nonpreferred in Group 5 is additional to this difference in information.

Since it has now been demonstrated that *S*s did select their best solutions with some degree of accuracy, the comparison with the standard condition of Group 1 becomes important. The means of Group 1, shown in Table 6, are higher for three of the five problems than those of the preferred solutions of Group 3, shown in Table 8, hence instructions to write many and select the best were not helpful. The means of the preferred solutions of Group 5 are better than the means of Group 1 for four of the five problems, but the differences are all small. It appears that instructions to write many solutions lower mean quality but with criteria information *S*s can select one of their productions which is about equal to the single ones written under standard conditions. These comparisons are summarized in Table 10.

The frequency of superior solutions among the preferred solutions for all problems was 29 for Group 3 and 42 for Group 5. The standard condition of Group 1 yielded 47. These frequencies are comparable since *S*s in Group 1 wrote one solution each and *S*s in Groups 3 and 5 chose one preferred solution each. By this measure

TABLE 10

MEAN RATING AND NUMBER OF SUPERIOR SOLUTIONS TO ALL PROBLEMS BY FIVE GROUPS

Group	Mean quality	Frequency of superior solutions
1	7.84	47
2	6.56	100
3		
Preferred	7.42	29
Nonpreferred	6.68	97
4	7.31	131
5		
Preferred	8.02	42
Nonpreferred	7.06	107

the standard condition is at least as good as any of the variations.

Thus one measure, mean quality of solutions identified as best by *S*s, gives a slight advantage to Group 5, but the other measure, number of superior solutions, gives a slight advantage to Group 1. Hence no stable advantage can be claimed for either condition. In general this analysis of accuracy of judgment leads to the conclusion that the judgment of *S*s was good but not good enough. Although *S*s of Group 5 had a wide range of solutions to choose from and information about the criteria that the expert raters used, their preferred solutions were not consistently better than the single solutions of Group 1.

Order of Production

To determine whether the superior solutions appeared early or late in the course of production, the sequences produced by the four multiple-solution groups, which varied in length from 2 to 22 solutions, were divided into first and last halves. A few superior solutions produced in the middle position of sequences of odd length were discarded. The results of these counts are shown in Table 11. For Plot Titles and Sentences there was a preponderance of superior solutions in the last half, but this was significant only for Sentences ($p < .05$). The difference in totals across all five problems was not significant.

TABLE 11

NUMBER OF SUPERIOR SOLUTIONS IN FIRST AND LAST PORTIONS OF PRODUCTION SEQUENCES

Problem	First	Last
Plot Titles	40	55
Table Titles	25	27
Conclusions	44	47
Sentences	43	64
Cartoon	60	59
Total	212	252

These totals were then analyzed by experimental groups. For Group 2 there were significantly more superior solutions in the last half ($p < .01$). The remaining totals were in that direction though not significant. Group 2, compared to the other multiple-solution groups, most closely conforms to the instructional conditions used by Gerlach, et al. (1964) and Parnes (1961) where a production-order effect was found with less complex problems. Similarly, Gerlach et al. did not find this effect with criteria-cued or evaluation instructions.

Individual Differences and Correlations

Since each *S* wrote solutions to five problems, there were 10 interproblem correlations for the quality of the single solutions and the number of superior solutions in Group 1, and in each of the four multiple-solution groups there were 10 interproblem correlations for number of solutions, mean quality, and number of superior solutions. Scores were also available for the CQT, taken at entrance. This test has three parts —verbal, general information, and numerical—but inspection of the correlations for the part scores did not reveal any relations not shown by the total score, hence only the correlations for the total CQT are mentioned here. Also available were scores on the Michigan State University (MSU) Reading Test, designed to measure ability of college freshmen to read textual materials in several academic areas. Since there were 40 *S*s in each group, correlations above .31 are significant at the .05 level.

In Group 1 the correlations for the quality rating of the solution between problems, and between problems and freshman tests were all positive but low, with a me-

dian of .16. For number of superior solutions the parallel correlations were also low, and some were negative, but when each *S*'s total number of superior solutions on all five problems was treated as a score, this total correlated .33 with the score on the MSU Reading Test and .45 with CQT. These correlations from the single-solution group are the easiest to interpret because each *S* probably did his best on his single opportunity. They indicate that the abilities required for these five problems in the standard condition have something in common with the abilities required for conventional tests of college aptitude.

In each of the multiple-solution groups, Groups 2–5, there were 10 correlations between problems in number of solutions produced. The 40 correlations for all four groups ranged from 0 to .70, but most were in the .40s and .50s, and the median was .45. This is evidence of the well-known factor of verbal productivity or meaningful fluency. The correlations with Reading and CQT were negligible.

The 40 corresponding correlations for mean quality rating ranged from −.33 to .40, with a median of .11, and the correlations with aptitude test scores were negligible. Instructions to write many solutions probably have different effects on the quality standards of different *S*s.

The 40 correlations for number of superior solutions ranged from −.26 to .53, with a median of .06. When each *S*'s total number of superior solutions on all five problems was treated as a score, the eight correlations between this total and aptitude test scores were all positive but low.

The most interesting correlations in the data for Groups 2–5 are those bearing on the old question of the relation between quantity and quality. Do *S*s who write the most solutions write the best? There were five correlations between number of solutions produced and mean quality of these solutions, one for each problem in each group, 20 for the four groups. With two exceptions these correlations were negative, ranging from −.53 to .13, with a median of −.27. Each *S*'s total number of solutions on the five problems was also correlated with the mean quality of all these solutions, yielding correlations for Groups 2–5, respectively, of −.13, −.32, −.47, and −.13. From these two ways of working up the data on 160 *S*s, it appears that those who wrote many solutions tended to write solutions of inferior quality. Thus these data on individual differences in quality agree with the data on differences between conditions. Groups that wrote many solutions wrote solutions of lower quality than groups that wrote only one, and, within the multiple-solution groups, those who wrote more solutions than average wrote solutions of less-than-average quality.

Similar correlations were computed between number of solutions produced by each *S* and number of superior solutions. These 20 correlations ranged from −.28 to .40, with a median of .09. They were more often positive than negative, but they were far smaller than the correlations reported elsewhere for simpler problems. For four groups writing unusual uses, Parnes and Meadow (1959) reported correlations between total number of solutions and number of superior solutions ranging from .64 to .81. Gerlach et al. (1964) confirmed this with a correlation of .78 for a similar task. In the present experiment, with more substantial problems, the difference between multiple-solution groups, writing about six solutions, and the single-solution group was large enough that the former groups produced more superior solutions. But within a multiple-solution group the variation in productivity was relatively smaller, and, furthermore, the advantage of large quantity was offset by the reduction in quality indicated by the preponderantly negative correlations mentioned above.

Discussion

The Nature of the Problems

Some of the differences between the results of this experiment and others on productive thinking are due to the nature of the problems. Tasks like writing unusual uses for a brick or uncommon associations are open to the criticism that they require only a superficial fluency. To avoid this

criticism the present experiment employed substantial problems that required *S* to integrate the material presented before constructing appropriate solutions. Variations in quality of solutions are due to comprehension and organization of the material as well as the production process per se.

Although the five problems were similar in form and yielded similar results in general, some differences can be noted. Many solutions could be written for all problems, but Table Titles required succinct comprehensive titles, and the number of these was small. Likewise Conclusions required inferences that correctly integrated the data of the chart, and the number of these was small. Comments of *S*s and review of the solutions to these two problems indicated that multiple-solution instructions led many *S*s to break up the problems and write titles and conclusions to portions of them. Such partial solutions received low ratings because comprehensiveness and integration were criteria for good solutions. Table 4 shows that 62 table titles and 23 conclusions were independently given maximal ratings by the two judges, however. Hence this analysis does not mean that these two problems were especially difficult but rather that each had only a small number of superior solutions which could not be much increased by emphasis on quantity. Although for the problems as a whole, the groups that wrote many solutions wrote more superior solutions, the difference comes largely from Plot Titles, Sentences, and Cartoon.

None of the problems had only one right answer but on the continuum from divergent to convergent thinking Table Titles and Conclusions would be located more toward the convergent end than the other three problems. More important, the above analysis offers a preliminary differentiation of solution processes at different locations on this continuum.

Quantity and Quality

The increase in number of superior solutions that comes from an increase in quantity of solutions is in agreement with the results of research on brainstorming by Parnes and others and extends the results to more substantial problems. The explanation need not be the same, however. Since the instructions for Group 2 said nothing about ignoring quality or postponing judgment, the pure quantity effect alone is sufficient to account for the difference in number of superior solutions, a ratio of more than two to one, over Group 1. Approximately the same ratio is reported when brainstorming instructions are compared with standard instructions (Meadow, Parnes, & Reese, 1959) and when students who have had a course in creative problem solving are compared with a control group (Parnes & Meadow, 1960). By "pure quantity effect" we mean an effect that operates over all degrees of quality, so that when more solutions are produced, more superior solutions are produced, and that this effect is produced by instructions that simply request many solutions. It should be noted, however, that the research on creative problem solving by Parnes and others compared multiple-solution groups under different instructions while the present comparison is between multiple-solution groups and single-solution groups.

Another effect of the emphasis on quantity, not emphasized in previous research, is the reduction in mean quality. The difference between the fifty-ninth and seventy-seventh percentiles is not trivial. The only other study using this measure that has been found (Weisskopf-Joelson & Eliseo, 1961) reported a drop in mean quality following instructions not to be critical. In the present study the drop in quality occurred on all five problems; it is safe to assume that it is a general phenomenon.

Variability and Production

The intraindividual variability in quality of solutions has not been previously reported. A quality dimension on which a sample of products from one *S* can be located with adequate reliability is not often available. It is possible that intraindividual variability in quality of solutions is larger for these complex problems than for the simpler tasks of other researchers, but no comparative data are available.

Whatever may be the mechanism for the production of solutions, it is no doubt a complicated one, so that the output varies on the quality dimension—and probably on other dimensions, such as length, clarity of handwriting, etc., not considered in these experiments. The semiinterquartile range of the ratings of the 200 single solutions produced by Group 1, computed from the distributions shown in Table 4 is about 3. Since the median of the ranges of the individual *S*s of the multiple-solution groups is about 7, the semiinterquartile range for individuals can be estimated as about 2. This last estimate is very rough, but it suggests that the variability within individuals does not differ greatly from the variability between individuals. To account for this variability there is no necessity at the present time, therefore, to assume any other production process for the individual drawing repeatedly upon his own resources than we assume for different individuals drawing once upon different resources.

The Contribution of Judgment

At this point the usefulness of the two measures, mean quality of all solutions and number of superior solutions, should be examined. It is hard to imagine any situation in which production of solutions completes the thinker's assignment. Usually the assignment is to contribute one solution. If the thinker produces only one, the quality of that one is of course the only measure to be considered. If he produces many, he usually eliminates all but one and submits that one. If he chooses at random from his total production, which he probably can do, the appropriate measure for research is the mean quality of all solutions. It has been seen in the present experiment that the highest mean was obtained by those who wrote only one solution. Hence, if random choice is to be the final operation, the best instructions would be to write one solution.

The other measure, number of superior solutions, is the one that has most often been used in the research situation, but it is hard to imagine any other situation in which expert judges would examine all of anyone's productions and separate the superior ones from the others. This is a measure of theoretical interest because it pertains to an intermediate operation, but the important overall question is whether the thinker can produce a superior solution *and* identify it as such in order to complete his assignment. The present study shows that college students can identify their best solutions with some accuracy. But the accuracy is not high. Even when they produce superior solutions and try to select their best, their choice is no better, in general, than the one solution produced by those instructed to write only one.

Since production is not improved any more by the procedures of other experiments than by the simple quantity instructions of the present experiment, improvement of the final evaluation of the solutions would appear to be the most promising step toward improvement of the overall performance. If quantity instructions increase the number of superior solutions and *S* can be trained to identify his best with high accuracy, then perhaps the one solution selected will be superior to the one produced by those who produce only one.

The assumption that *S* produces a variety of responses one of which is selected by reinforcement contingencies in the environment is a familiar one, under such names as trial and error, but the selection of one response by *S* as a separate operation has been generally overlooked.

Experiment II: Evaluation of Judgment Training

There are two questions about improvement of productive thinking by training in judgment. (*a*) Under multiple-solution instructions, will judgment training increase the difference between preferred and nonpreferred solutions relative to the difference obtained without training? (*b*) Will the single preferred solutions, selected by *S* after judgment training, be superior to the single solutions produced under standard single-solution conditions?

The literature offers little guidance in the development of training programs for judgment of solutions to problems. The one relevant experiment (Johnson & Zerbolio, 1964) gave brief practice in judging

plot titles, but only a small, nonsignificant improvement was obtained. Hence three variations of materials and procedures were tried out in a pilot study with 111 *S*s judging solutions to the sentences problem and 104 *S*s judging solutions to the conclusions problem. Observation of the behavior of *S*s, as well as the objective results, led to preparation of materials in booklets that seemed to combine the advantages of all three variations.

Sentences and Conclusions were chosen because the results of Experiment I indicated that these two were the most dissimilar. At this point there is more to be learned about possibilities of improvement from these than from two that are more similar.

Method

Training Materials

The solutions were taken from the results of previous experiments; only those on which the two expert judges agreed exactly were used. Rating guides, which the expert judges had written for their own use, were given to *S*s. Practice A displayed four poor, four good, and four superior solutions, with an explanation for the rating of each and appropriate reference to the rating guide. The *S*s were told to study these and, when they understood the characteristics of superior solutions, to try Practice B.

Practice B presented 14 triads of solutions to the Sentences problem (or 13 to Conclusions), with instructions to select the best one of each. Official judgments of the first two triads, with explanations, were given on the next page, and official judgments of the remainder were also given to permit each *S* to calculate a score for his accuracy of judgment. The first few judgments were made rather easy by grouping one very good solution with two poor ones.

Practice C stressed the contrast of superior and inferior solutions. Seven of each were printed, with official judgments and explanations for them. Then appeared a block of seven superior solutions with blank lines for *S* to write the characteristics of superior solutions. This was followed by a block of seven inferior solutions with lines for the characteristics of inferior solutions. This format was repeated with six more superior and inferior solutions.

Procedure

The sequence of events for the training groups was: writing solutions to a problem, practice in judging, return to the solutions first written to select the best. The instructions for the last phase were as follows:

Up to now you have been practicing the evaluation of sentences (or conclusions) which were written by *other* students. Now the real test of your judgment ability comes. Go back to the sentences which you wrote on the first page. Read each sentence carefully and select the three (3) best according to the criteria of superior sentences which you have been using. Put an "X" by each of these three.

Then, reread these three sentences and put a line completely around the *one* best.

Three training groups had the same materials but in one condition, called individual practice, each *S* worked alone, as under classroom or examination conditions.

The model for dyadic practice was the common procedure in research teams when two investigators practice together to work up good interjudge agreement in rating statements of opinion, taped interviews, and solutions to problems. The advantages of dyadic practice seem to be motivational, in that the two judges stimulate each other, and also intellectual, since in discussing disagreements they are likely to read the solutions and the rating guide more carefully and thus to learn the criteria of judgment more specifically. The pairs were assembled in small sections of 8–12 so that *E* had only 4–6 pairs to supervise. The *S*s entered a room and took seats at tables for two, hence the pairing was presumably random, although it is possible that friends might take adjacent seats and thus find themselves working as partners. Those dyads that were slow in initiating conversation were encouraged by questions and procedural, but not substantive, hints from *E*.

The third training condition was tutorial practice, one *S* working with one *E*. Because of the time required this condition was limited to one problem, Conclusions. The *E* gave direct help in clarifying criteria, explaining disagreements, and interpreting the rating guide.

Each of the training procedures required 40–60 minutes. Groups that wrote many solutions and selected their best were allowed 7 minutes per problem. Those who wrote only one solution required only a few minutes per problem.

Subjects

The *S*s were students in elementary psychology classes at Michigan State University, 20 in each condition.

Group 1 was a control group, instructed to write one solution. Since little time was required, these *S*s were given both Sentences and Conclusions, half in each order.

Group 2 was instructed to write one solution, but was given the criteria for good solutions, as in Experiment I. Since the criteria for good solutions were given to training groups, a single-solution group given the criteria was necessary for comparison. These *S*s were given both problems, half in each order.

Group 3 wrote many solutions and then selected their best. The criteria for good solutions were given. Since relatively little time was required, these *S*s were given both problems, half in each order.

Group 4 had individual practice, with one problem.

Group 5 had dyadic practice, with one problem.

Group 6 had tutorial practice, with Conclusions only.

When *S*s were assembled for dyadic practice, one or two were withdrawn to another room for tutorial practice, leaving an even number.

Since sex differences were not considered in Experiment I but are occasionally reported in problem-solving research, these groups were roughly balanced for sex.

Results

The solutions were processed and rated as described for Experiment I, except as noted below. Interjudge agreement was adequate. Order effects were possible in Groups 1, 2, and 3 but, in respect to mean quality of solutions, these were small and inconsistent. Sex differences were also small and inconsistent in all groups. Hence order and sex have been ignored in subsequent analyses.

Sentences

Mean ratings of the solutions produced by the two single-solution groups were about the same; the criteria cues were not effective (see Table 12). As in Experiment I the multiple-solution groups wrote varying numbers of solutions, averaging about six or seven per *S*, and the mean quality of these solutions was lower than for single-solution groups.

The first question about the effects of judgment training is whether the difference of the nonpreferred solutions is larger for groups with training than for Group 3,

TABLE 12
MEAN QUALITY OF SENTENCES PRODUCED AND JUDGED UNDER FIVE CONDITIONS

Condition	Total	Preferred	Nonpreferred	Diff.	t
One solution	9.15				
One solution, criteria	9.00				
Many solutions	8.47	8.75	8.29	.46	.85
Many solutions, individual training	8.12	9.35	7.85	1.50	3.84*
Many solutions, dyadic training	8.75	10.45	8.31	2.14	4.28*

* $p < .01$.

without training. Table 12 shows larger differences in both training groups. To test for significance a difference score was computed for each *S*, as in Experiment I, and an analysis of variance showed a significant variance between groups ($F = 3.25$; $p < .05$). Comparing the two training groups with Group 3, as a control group, by Dunnett's t statistic, Group 5 (with dyadic practice) was superior to Group 3 ($p < .025$). In general it appears that the judgment practice, especially practice in pairs, was effective in helping *S*s select their best solutions.

The other question about the effectiveness of judgment training involves a comparison with the standard condition in which *S* simply writes one solution. Since each *S* in Groups 1 and 2 wrote only one solution and each *S* in Groups 3, 4, and 5 selected one preferred solution, there were 20 solutions in each of the five groups for this comparison. Analysis of variance disclosed a difference between groups at the .05 level. Groups 1 and 2 may be combined to represent the standard condition or control group, and in comparison with this group of 40 solutions the solutions of Group 5 were superior by the Duncan multiple range test adapted for unequal n's ($p < .05$). Neither Group 3 nor Group 4 was significantly different from the standard condition by this measure.

Some of these results are displayed graphically in Figure 2.

Analysis of the frequency of superior solutions shows the same relations but the numbers are small. To approximate the top 10% of the 434 solutions the nearest cutting score was 11, and there were 38 ratings of 12, 13, and 14. The two single-solution groups produced 3 each, while Groups 3, 4, and 5 produced 11, 7, and 14, respectively. As to the effects of training, there were only 2 superior solutions among the preferred solutions of Group 3, but Group 4 chose 3 and Group 5 chose 7. Thus only Group 5, with dyadic practice, was better than the control groups.

As a measure of the effects of training, the frequency of superior solutions is somewhat misleading because some *S*s wrote

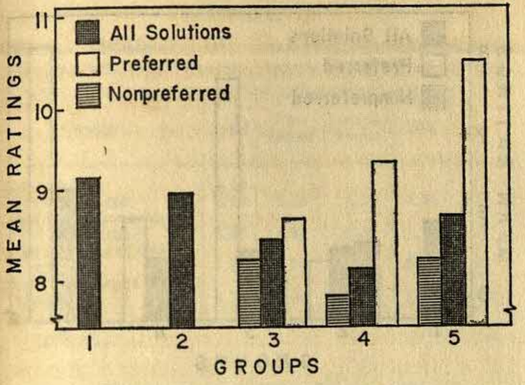


FIG. 2. Effects of judgment training on solutions to sentences problem. (The *S*s in Groups 1 and 2 wrote one solution each, but Group 2 had the criteria for good solutions. Groups 3, 4, and 5 wrote many solutions, but Group 4 had individual-judgment training and Group 5 had dyadic-judgment training.)

more than one but could not select more than one as the best. If only the number of *S*s who produced at least one superior solution are considered, the percentages of preferred solutions relative to these were 60 for Group 4, and 70 for Group 5, as contrasted with only 33 for the untrained Group 3. These data are too meagre to stand by themselves but they agree, by another method of analysis, with the data on mean quality.

Conclusions

Mean quality of the solutions written by Group 2, having the criteria for good solutions, was slightly higher than that for Group 1, but the difference was not significant (see Table 13). The multiple-solution

TABLE 13

MEAN QUALITY OF CONCLUSIONS PRODUCED AND JUDGED UNDER SIX CONDITIONS

Condition	Total	Preferred	Nonpreferred	Diff.	t
One solution	8.20				
One solution, criteria	9.50				
Many solutions	7.07	8.65	6.49	2.16	2.40*
Many solutions, individual training	5.87	7.70	5.32	2.38	2.66**
Many solutions, dyadic training	6.38	8.30	5.81	2.49	2.78**
Many solutions, tutorial training	5.80	6.90	4.98	1.92	1.70

* $p < .05$.
** $p < .01$.

groups wrote 4–6 solutions per *S*, of generally lower quality, as in Experiment I. Figure 3 displays these data graphically.

In three multiple-solution groups the preferred solutions were significantly superior to the nonpreferred solutions. Evidently judgment was easy with the criteria of a good solution, and judgment training did not add much to accuracy of judgment. Hence the differences between preferred and nonpreferred solutions were about the same in all four groups.

The quality of the preferred solutions is the joint effect of the quality of the solutions produced and accuracy in selecting the best. Table 13 shows considerable variation

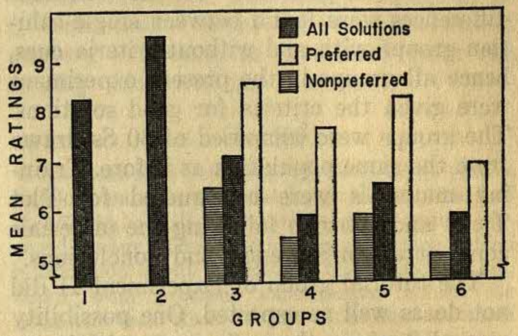


FIG. 3. Effects of judgment training on solutions to conclusions problem. (The *S*s in Groups 1 and 2 wrote one solution each, but Group 2 had the criteria for good solutions. Groups 3–6 wrote many solutions, but Group 4 had individual-judgment training, Group 5 had dyadic training, and Group 6 had tutorial training.)

in quality of solutions produced by the four multiple-solution groups. Group 3, which produced solutions of relatively high quality, had preferred solutions of high quality, although not as high as the single solutions of Group 2. Since Group 6 produced solutions of low quality, this experiment may not provide a fair test of the value of tutorial training.

Analysis of the number of superior solutions disclosed the same relations by another measure.

The two problems chosen for Experiment II were the most diverse of the five problems of Experiment I. Conclusions was the only problem on which multiple-solution groups did not produce distinctly more superior

solutions than single-solution groups (see Table 7). Conclusions may be a problem in which accuracy of judgment of solutions is so good that there is little room for improvement. Sentences, on the other hand, were judged poorly without special training. Obviously, what is needed is a repetition of Experiment II with other problems.

EXPERIMENT III: EVALUATION OF JUDGMENT TRAINING

The plot-title problem and the cartoon problem were chosen for further evaluation of judgment training, and the same procedures were followed, with a few exceptions. In Experiment II no clearcut differences were found between single-solution groups with and without criteria cues, hence all groups in the present experiment were given the criteria for good solutions. The groups were composed of 30 *S*s drawn from the same population as before. Training materials were constructed for Plot Titles and Cartoon following the materials constructed for Sentences and Conclusions.

The tutorial group of Experiment II did not do as well as expected. One possibility was that since these *S*s were withdrawn from one room and taken to another, the special conditions may have lowered their productivity prior to judgment training. Hence, in the present experiment this group wrote their solutions in the same room with the other *S*s and were withdrawn later for judgment training.

Results

Plot Titles

The results, analyzed as before, are shown in Table 14 and Figure 4. The in-

TABLE 14

MEAN QUALITY OF PLOT TITLES PRODUCED AND JUDGED UNDER FIVE CONDITIONS

Condition	Total	Preferred	Nonpreferred	Diff.	t
One solution	5.97				
Many solutions	5.78	5.60	5.86	−.26	.89
Many solutions, individual training	5.77	77.0	5.33	2.37	4.32*
Many solutions, dyadic training	5.46	5.97	5.40	.57	1.44
Many solutions, tutorial training	6.37	8.13	5.88	2.25	4.67*

* $p < .01$.

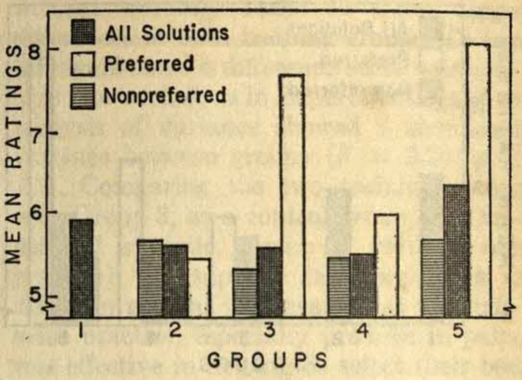


FIG. 4. Effects of judgment training on solutions to plot-title problem. (The *S*s in Group 1 wrote one solution each, with the aid of criteria cues. Groups 2–5 wrote many solutions, but Group 3 had individual-judgment training, Group 4 had dyadic training, and Group 5 had tutorial training.)

dividual training and the tutorial training yielded significant differences but the difference due to dyadic training was surprisingly small. Analysis of variance of the preferred solutions from the multiple-solution groups together with the single solutions from Group 1 disclosed a significant between-groups difference ($p < .01$). As compared with Group 1, the preferred solutions of Group 3 were better ($p < .05$) and so were those of Group 5 ($p < .01$).

Table 14 shows that Group 5 wrote solutions of a higher mean quality than Group 1. This small difference, which is not significant, is the only case in which instructions to write many solutions did not cause a drop in mean quality. In Experiment I it was Plot Titles that showed the smallest drop in mean quality as a consequence of multiple-solution instructions.

Cartoon

The results for the cartoon problem are shown in Table 15 and Figure 5. All three types of judgment training resulted in significant differences between preferred and nonpreferred solutions, while the corresponding difference for Group 2, without training, was negligible. The preferred solutions of the groups with judgment training were approximately of the same mean quality as the single solutions of Group 1.

TABLE 15

Mean Quality of Cartoon Completions Produced and Judged under Five Conditions

Condition	Total	Preferred	Nonpreferred	Diff.	t
One solution	8.14				
Many solutions	6.64	6.46	6.74	−.28	1.12
Many solutions, individual training	6.35	8.14	6.17	1.97	3.60*
Many solutions, dyadic training	6.56	8.74	5.86	2.88	3.60*
Many solutions, tutorial training	6.73	8.04	6.61	1.43	3.59*

* $p < .01$.

Experiment IV: Judgment Without Production

In Experiments II and III judgment of the solutions was complicated by production of the solutions. This condition is representative of common production-and-judgment situations and is a necessary complication for some experiments, but certain comparisons can be made more clearly by isolating the judgment process. One question concerns the effects of criteria cues on judgment. Since criteria cues improved production in the multiple-solution conditions of Experiment I, it is possible that information about criteria was responsible for a major part of the improvement attributed to judgment training in Experiments II and III. This hypothesis is most plausible for Conclusions because

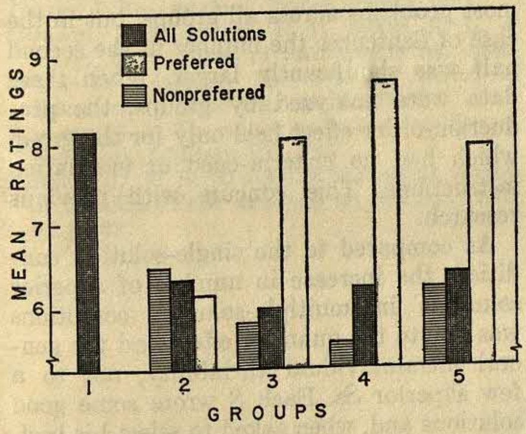


Fig. 5. Effects of judgment training on solutions to cartoon problem. (The *Ss* in Group 1 wrote one solution each, with the aid of criteria cues. Groups 2–5 wrote many solutions, but Group 3 had individual-judgment training, Group 4 had dyadic training, and Group 5 had tutorial training.)

of the small number of good solutions. The alternative hypothesis is that judgment training is necessary. Perhaps information about criterial and irrelevant dimensions can be effectively utilized only after specific training, with appropriate feedback.

The pilot study, mentioned as a preliminary to Experiment II, included a multiple-choice judgment test. Although this was intended more for training than for evaluation, the data suggested possible use as a dependent variable. Therefore new multiple-choice judgment tests were constructed for Conclusions, Sentences, and Cartoon, incorporating solutions produced by *Ss* of previous experiments and rated by the expert judges. The tests consisted of 10 five-alternative items for each problem of varying difficulty, as estimated from the quality ratings of the solutions.

Method

The standard condition required *S* to read a problem and then take the judgment test. Each of the 42 *Ss* had all three problems, in counterbalanced orders. The criteria-cued condition added information about the criteria of a good solution, taken from previous experiments, before the judgment test. Each of the 41 *Ss* had all three problems, in counterbalanced orders. The judgment training required more time, hence each *S* had only one problem, then the training workbook, and then the judgment test. Thus there were 126 *Ss* in this condition, 42 for each problem. This training corresponds to the individual training of Experiments II and III. In all conditions, however, the final judgments on the test would not be considered ego-involving, since *Ss* were judging solutions written by others.

Booklets were assembled and distributed to small groups of volunteer *Ss* who came to the research building to participate in productive thinking research and receive class credit.

Results

Keys for the multiple-choice judgment test were made up in advance, and the tests were scored mechanically. Mean scores are shown in Table 16. The standard deviations for the nine distributions were quite similar, varying from 1.23 to 1.66. Analysis of variance yielded significant F ratios for each problem, and for each problem the mean after judgment training was different from each of the other two means ($p < .01$), according to the Duncan multiple range test. In general, these results support

TABLE 16

MEAN SCORES ON JUDGMENT TEST UNDER THREE CONDITIONS OF INSTRUCTION AND TRAINING

Problem	Standard condition	Criteria cues	Judgment training	*F*
Conclusions	3.76	3.95	5.17	10.33*
Sentences	2.69	2.55	4.88	41.27*
Cartoon	3.86	4.48	6.17	19.84*

* $p < .001$.

the effectiveness of judgment training even when criteria cues have been supplied.

CONCLUSIONS

A few methodological conclusions are in order. These experiments demonstrate the potentialities of separating production from judgment and manipulating each process separately. College students can follow simple instructions about production and judgment without difficulty. Multiple-choice tests of judgment have certain advantages also; judgment without production can be studied as well as production without judgment.

Despite the high agreement between the two raters in judging *Ss'* solutions, the use of ratings always raises methodological suspicions because two judges can agree on many dimensions, not all of which are pertinent to the hypotheses under consideration. It is comforting to note that Ss' preferred solutions got higher ratings from the expert judges than the nonpreferred solutions, and that the difference was enhanced by information about the the judges' criteria and by practice in judgment.

It is obvious, when overall performance is considered, that the ordinary single-solution condition should be included. Much more enthusiastic conclusions about specific procedures could be written if this condition had been left out. The standard procedure of asking *S* to write a solution to a problem is a hard one to beat.

Instructing *S* to select his best solution is a useful research procedure. Having *S* produce many solutions illuminates the production process, but it is seldom an end in itself. Someone has to select the best solutions.

The value of including several problems was amply supported. The five problems were similar in certain aspects of the results, but different in other aspects.

As compared with the standard condition of writing one solution, instructions to write many solutions yielded large effects. Mean quality went down and the number of superior solutions went up. Such instructions may be augmented by instructions to defer judgment or ignore quality, but the addition of this feature was not studied here. Instructions simply to write many solutions resulted in roughly double the number of superior solutions, about the same increase as reported for more complicated instructions. These instructions may be augmented in the other direction also, by including the criteria for good solutions. This addition decreased number produced and increased both mean quality and number of superior solutions. The reduction in quality with instructions to increase quantity is consonant with the negative correlation for individual differences in quality and quantity.

When instructed to write many solutions, each *S* wrote solutions of a wide range of quality. The number of superior solutions was about the same in the first and second halves of a 7-minute production period for most problems across all groups, but in the case of Sentences the number in the second half was significantly larger. When these data were analyzed by groups, the production-order effect held only for the group which had no criteria-cued or evaluation instructions. This concurs with previous research.

As compared to the single-solution condition, the increase in number of superior solutions in multiple-solution conditions was due to the quantity effect and the general intraindividual variability, not to a few superior *Ss*. Each *S* wrote some good solutions and, when asked to select his best, selected one that was likely to be better than the others. Even so, the preferred solutions were seldom better than the solutions of those who wrote only one each.

Attempts to improve overall performance by three types of judgment training—individual, dyadic, and tutorial—were generally successful in that the differences between preferred and nonpreferred solutions were increased, though this improvement varied across problems and type of training. Dyadic training, for example, was particularly successful on Sentences and Cartoon but not on Plot Titles. When conditions were favorable, that is, when instructions to write many solutions did not reduce quality severely and when the training in judgment was quite successful, the preferred solutions of the multiple-solution groups were superior to the solutions written by the standard single-solution groups. And a control experiment with a multiple-choice test of judgment demonstrated that the improvement was due to the judgment training, not merely to information about the criteria for good solutions.

REFERENCES

BLOOM, B. S., & BRODER, L. J. Problem-solving processes of college students. *Supplementary Educational Monographs,* 1950, No. 73.

BURAK, B. The nature and efficacy of methods of attack on reasoning problems. *Psychological Monographs,* 1950, **64** (7, Whole No. 313).

CHRISTENSEN, P. R., GUILFORD, J. P., & WILSON, R. C. Relations of creative responses to working time and instructions. *Journal of Experimental Psychology,* 1957, **53,** 82–88.

DEWEY, J. *How we think.* Boston: Heath, 1910.

GERLACH, V. S., SCHUTZ, R. E., BAKER, R. L., & MAZER, G. E. Effects of variations in test directions on originality test response. *Journal of Educational Psychology,* 1964, **55,** 79–83.

GUILFORD, J. P. *The nature of human intelligence.* New York: McGraw-Hill, 1967.

JOHNSON, D. M. Serial analysis of thinking. *Annals of the New York Academy of Sciences,* 1960, **91,** 66–75.

JOHNSON, D. M. Formulation and reformulation of figure-concepts. *American Journal of Psychology,* 1961, **64,** 418–424.

JOHNSON, D. M., & JENNINGS, J. W. Serial analysis of three problem-solving processes. *Journal of Psychology,* 1963, **56,** 43–52.

JOHNSON, D. M., & ZERBOLIO, D. J. Relations between production and judgment of plot-titles. *American Journal of Psychology,* 1964, **77,** 99–105.

MEADOW, A., & PARNES, S. J. Evaluation of training in creative problem solving. *Journal of Applied Psychology,* 1959, **43,** 189–194.

MEADOW, A., PARNES, S. J., & REESE, H. Influence of brainstorming instructions and problems sequence on a creative problem solving test. *Journal of Applied Psychology,* 1959, **43,** 413–416.

NEWELL, A., SHAW, J. C., & SIMON, H. A. Elements of a theory of human problem solving. *Psychological Review,* 1958, **65,** 151–166.

PARNES, S. J. Effects of extended effort in creative problem solving. *Journal of Educational Psycology,* 1961, **52,** 117–122.

PARNES, S. J., & MEADOW, A. Effects of brainstorming instructions on creative problem solving by trained and untrained subjects. *Journal of Educational Psychology,* 1959, **50,** 171–176.

PARNES, S. J., & MEADOW, A. Evaluation of persistence of effects produced by a creative problem-solving course. *Psychological Reports,* 1960, **7,** 357–361.

PATRICK, C. Creative thought in poets. *Archives of Psychology,* 1935, No. 178.

PATRICK, C. Creative thought in artists. *Journal of Psychology,* 1937, **4,** 35–73.

WALLACH, M. A., & KOGAN, N. *Modes of thinking in young children.* New York: Holt, Rinehart & Winston, 1965.

WALLAS, G. *The art of thought.* New York: Harcourt, Brace, 1926.

WEISSKOPF-JOELSON, E., & ELISEO, T. S. An experimental study of the effectiveness of brainstorming. *Journal of Applied Psychology,* 1961, **45,** 45–49.

WILSON, R. C., GUILFORD, J. P., & CHRISTENSEN, P. R. The measurement of individual differences in originality. *Psychological Bulletin,* 1953, **50,** 363–370.

(Received May 17, 1968)

EDUCATION ... PSYCHOLOGICAL